Elements of Statistical Learning
Chapter 2
Overview of Supervised Learning

Keng-Chi Chang

2018-01-02

Topics

- Statistical Decision Theory (2.4)
- Special Cases: Linear Regression vs. k-Nearest Neighbors (2.3)
- Bias-Variance Decomposition (2.9)
- Curse of Dimensionality (2.5)
- Classes of Restricted Estimators (2.6–2.8)

Other References

- Chapters 2 and 3 of *An Introduction to Statistical Learning*
- Ryan Tibshirani's Lecture Note

Goal: Predict Y from X

- Let $(X, Y) \sim$ iid $P(X, Y)$
- Come up with function f to predict X from Y
- (Squared Error) Loss $L(Y, f(X)) = (Y - f(X))^2$
- Expected loss $E[(Y - f(X))^2]$ is minimized at $f(x) = E[Y|X = x]$
  - Shown by subtracting and adding $E[Y|X]$
- $f(x) = E[Y|X = x]$ is called the true regression function

Use Data to Estimate f

- Now we have (training) data

$$D_n = \{(x_1, y_1), \cdots, (x_n, y_n) \,|\, x_i \in \mathbb{R}^p\} \sim \text{ iid } P(X, Y)$$

- Write $X = (x_1^T, \cdots, x_n^T)^T$, $Y = (y_1, \cdots, y_n)^T$
- Use D_n to construct $\hat{f}$ that mimics f
- That is, we want to find $\hat{f}$ with smaller $E[(Y - \hat{f}(X))^2]$
- Notice: $\hat{f}$ depends on data X, hence random

Use Data to Estimate f

- What we can observe is training error

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

- This is a sample analog of prediction error *conditional on the (training) data*

$$E[(Y - \hat{f}(X))^2 \mid D_n]$$

- Any function $\hat{f}$ passes through D_n can be a minimizing solution
- Furthermore, what we really want to know is the *unconditional* prediction error that considers data outside D_n

$$E[(Y - \hat{f}(X))^2] = E[E[(Y - \hat{f}(X))^2 \mid D_n]]$$

Linear Regression

- If we believe the relationship between X and Y is close to linear, can make this functional form assumption on f and assume

$$f(X) = E[Y|X] \approx X\hat{\beta} = \hat{f}(X)$$

- This reduces model complexity (by assuming functional form)
- We only need to estimate a finite number of parameters β
- This is a parametric approach to model $f(X)$

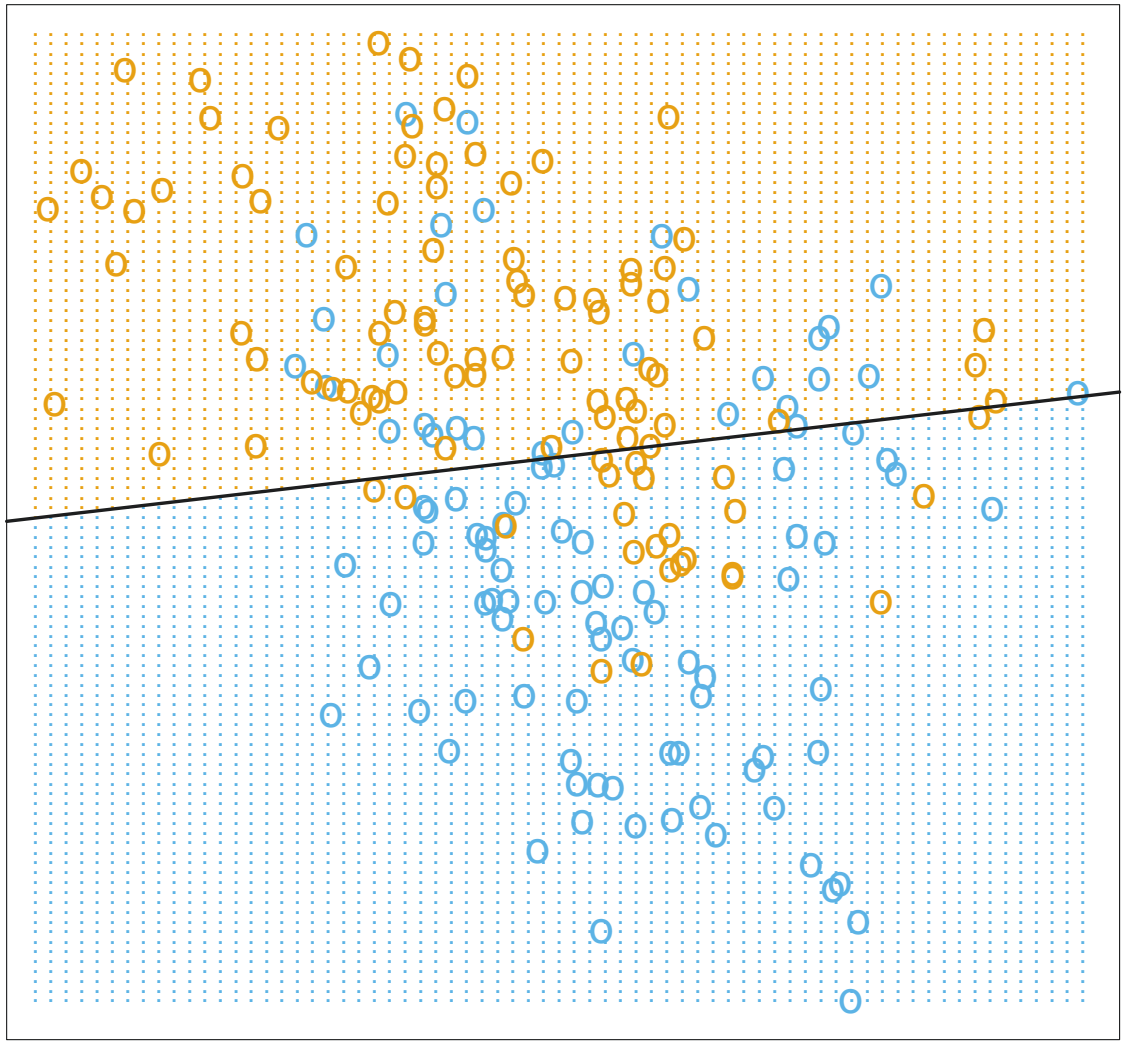


FIGURE 2.1. *A classification example in two dimensions. The classes are coded as a binary variable (*BLUE $= 0$, ORANGE $= 1$*), and then fit by linear regression. The line is the decision boundary defined by* $x^T \hat{\beta} = 0.5$*. The orange shaded region denotes that part of input space classified as* ORANGE*, while the blue region is classified as* BLUE*.*

k-Nearest Neighbors

- One may not want to make functional form assumptions of f
- One idea: A sample analog of $f(x) = E[Y|X = x]$ can be

$$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

  where N_k is the set contains all k nearest neighbors of x
- This is a more flexible class of function
- But also comes with costs, since this is similar to estimating infinite dimensions of parameters
- This is a non-parametric approach to model $f(X)$

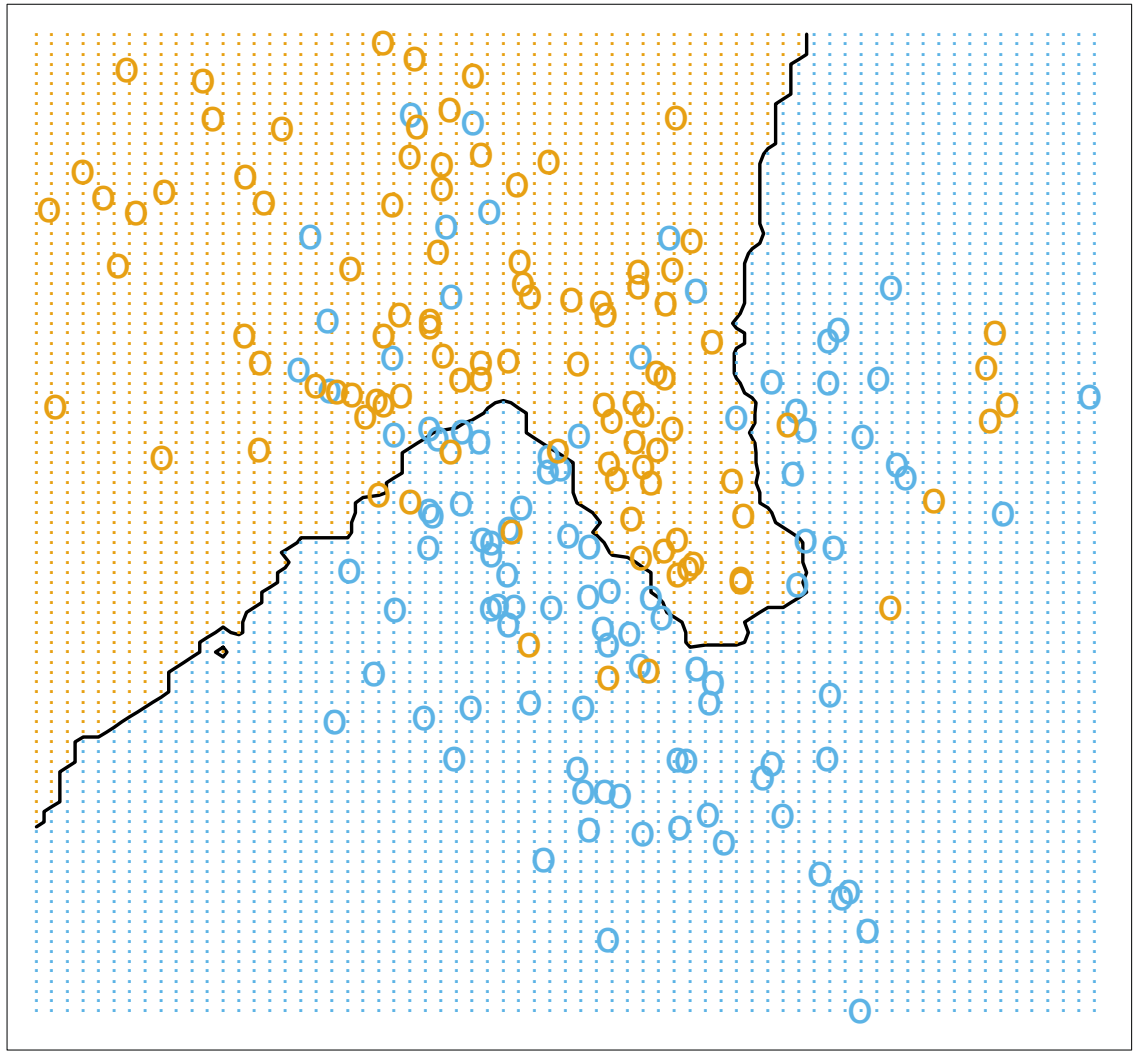


FIGURE 2.2. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$) *and then fit by* 15*-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the* 15*-nearest neighbors.*

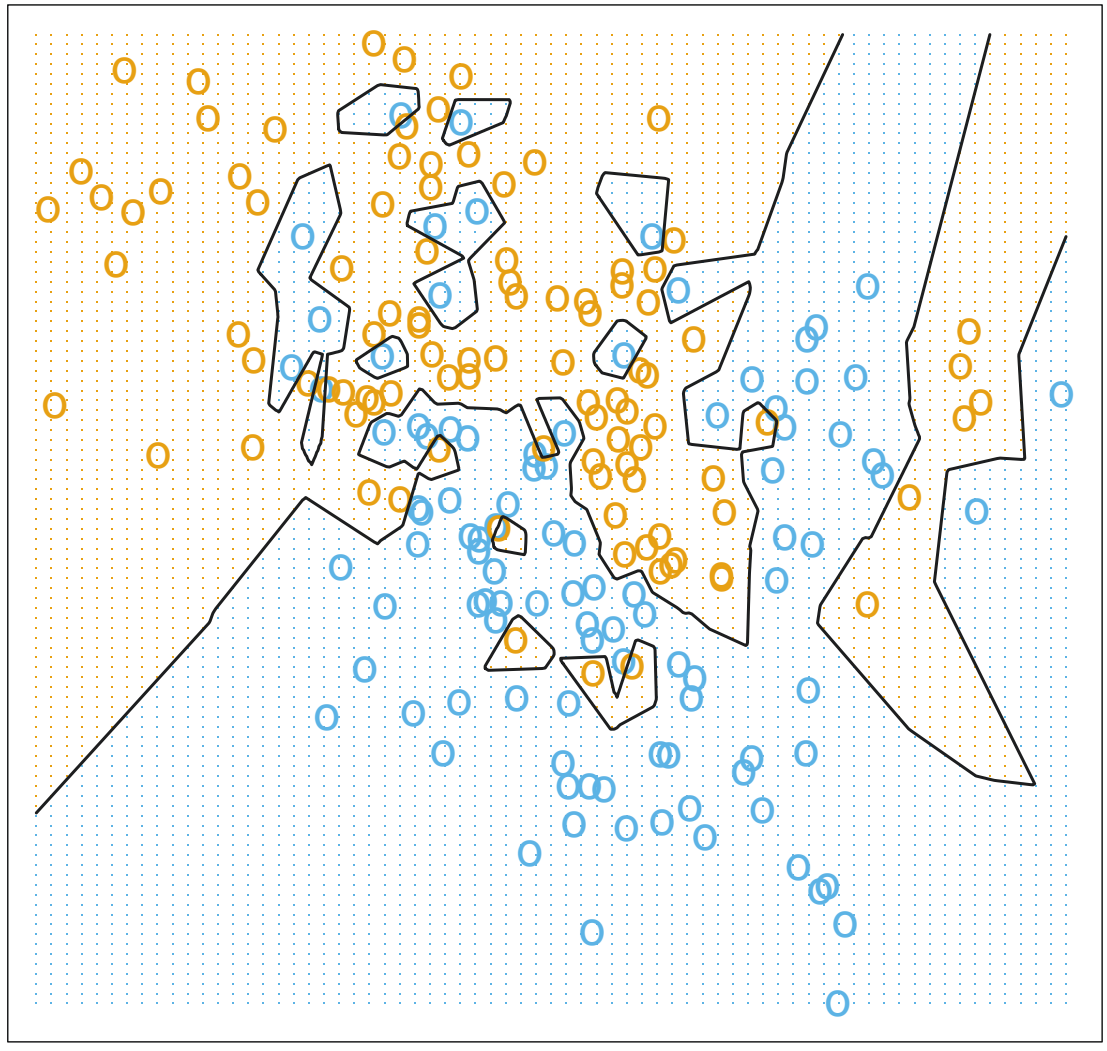


FIGURE 2.3. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$)*, and then predicted by* 1*-nearest-neighbor classification.*

Irreducible and Mean Squared Errors

- Suppose $Y = f(X) + e$ where $e \sim (0, \sigma^2)$ and $e \perp X$
- Conditional on $X = x$, subtracting and adding $f(x)$ from the prediction error, we can write

$$E[(Y - \hat{f}(X))^2 \mid X = x] = E[(Y - f(x))^2] - E[(f(x) - \hat{f}(x))^2]$$

  since $E[e] = 0$ and $e \perp X$ by construction
- $E[(Y - f(x))^2] = \sigma^2$ is irreducible, even if $\hat{f} = f$
- $E[(f(x) - \hat{f}(x))^2] = \text{MSE}(\hat{f}(x))$ is the mean squared error of the function estimator $\hat{f}(x)$

Bias-Variance Decomposition

- Further decompose MSE by subtracting and adding $E[\hat{f}(x)]$

$$\begin{aligned} E[(f(x) - \hat{f}(x))^2] &= E[(f(x) - E[\hat{f}(x)])^2] + E[(E[\hat{f}(x)] - \hat{f}(x))^2] \\ &= (f(x) - E[\hat{f}(x)])^2 + E[(\hat{f}(x) - E[\hat{f}(x)])^2] \\ &= \text{Bias}^2(\hat{f}(x)) + \text{Var}(\hat{f}(x)) \end{aligned}$$

  where the first equality is due to $E[E[\hat{f}(x)] - \hat{f}(x)] = 0$

- The unconditional prediction error is thus

$$\begin{aligned} E[(Y - \hat{f}(X))^2] &= E_X[E[(Y - \hat{f}(X))^2 | X]] \\ &= \sigma^2 + E_X[\text{Bias}^2(\hat{f}(X))] + E_X[\text{Var}(\hat{f}(X))] \end{aligned}$$

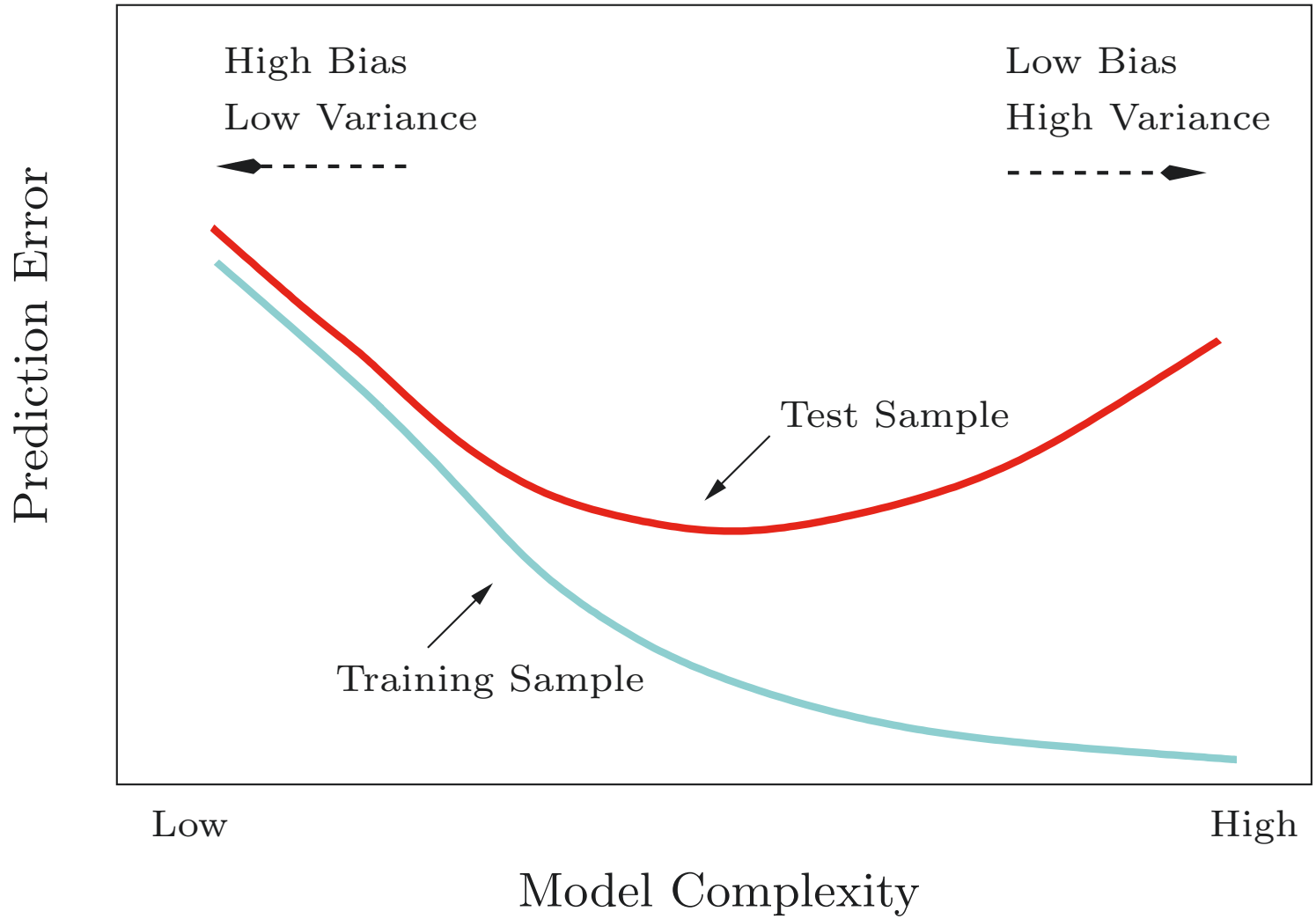


FIGURE 2.11. *Test and training error as a function of model complexity.*

Linear vs. KNN Regressions

- Bias: How expected $\hat{f}$ is close to f $\quad f(x) - E[\hat{f}(x)])$
- Variance: How variable is $\hat{f}$ itself $\quad E[(\hat{f}(x) - E[\hat{f}(x)])^2]$
- Linear Regression: If $f(X) = X\beta$, then $E[\hat{\beta}] = \beta$, and

$$MSE(\hat{f}(x)) = 0 + (x^T(X^TX)^{-1}x)\sigma^2$$

  - If f is linear, then unbiased; if not, possibly (high) bias

- k-Nearest Neighbors:

$$MSE(\hat{f}(x)) = \left(f(x) - \frac{1}{k}\sum_{x_i \in N_k(x)} f(x_i)\right)^2 + \frac{\sigma^2}{k}$$

  - More flexible, can lower bias, but also increase variability

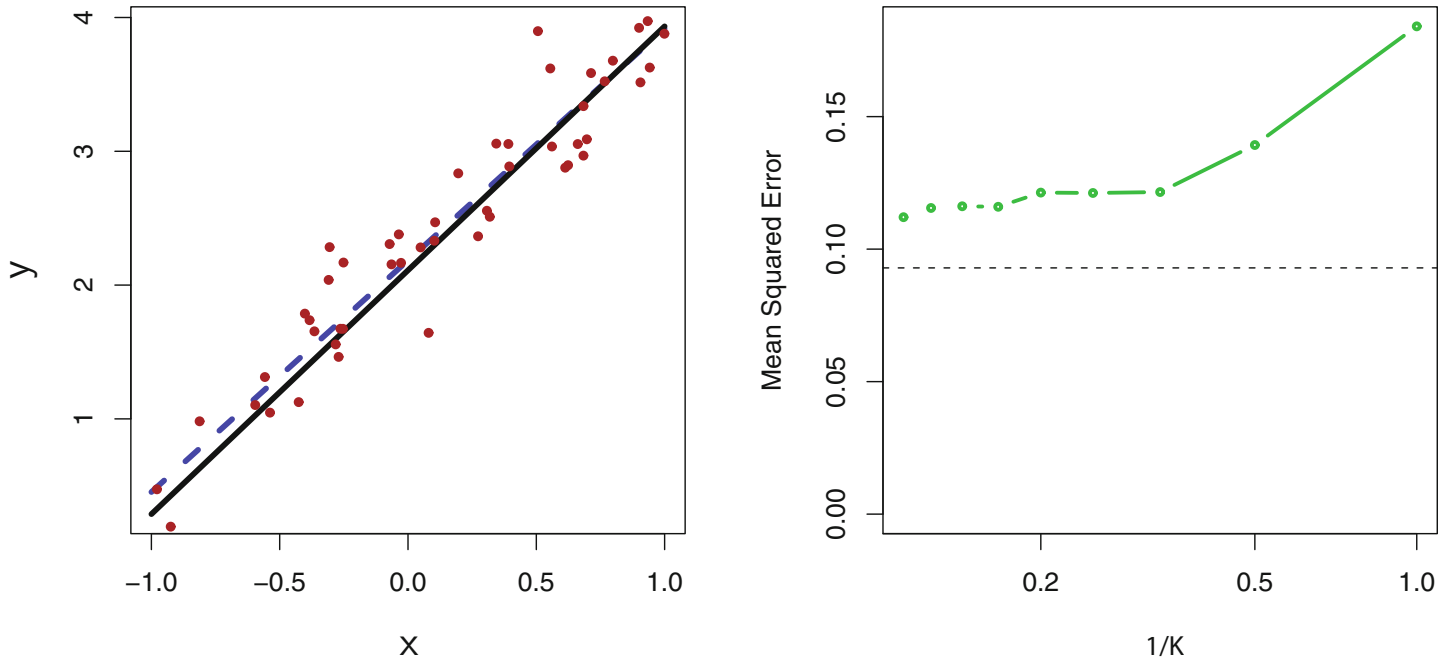
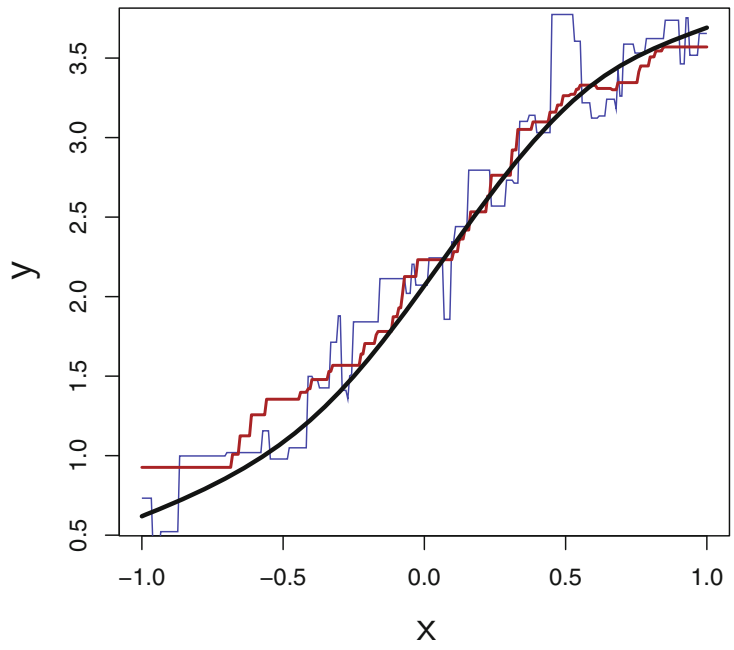
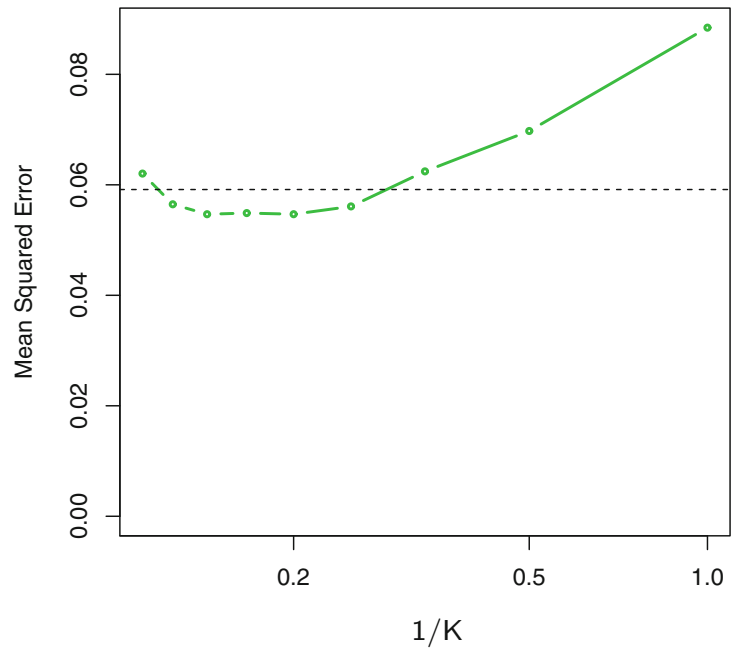
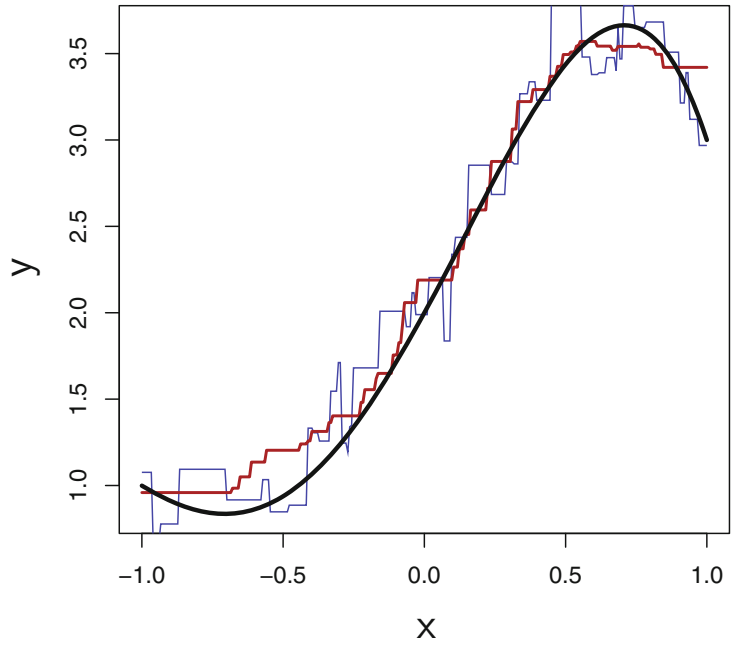
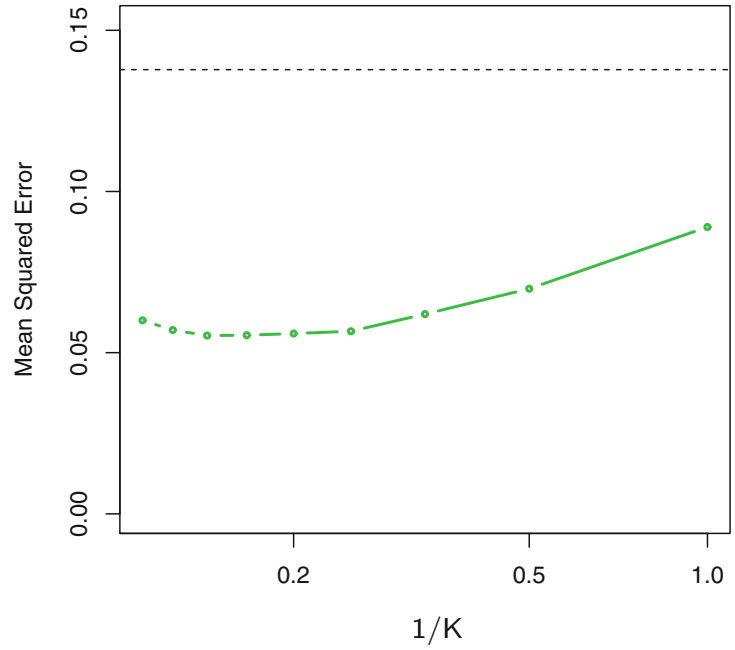


FIGURE 3.18. *The same data set shown in Figure 3.17 is investigated further.* Left: *The blue dashed line is the least squares fit to the data. Since $f(X)$ is in fact linear (displayed as the black line), the least squares regression line provides a very good estimate of $f(X)$.* Right: *The dashed 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 $1/K$.*


y
X
0.5
1.0
1.5
2.0
2.5
3.0
3.5
−1.0
−0.5
0.0
0.5
1.0


Mean Squared Error
1/K
0.00
0.02
0.04
0.06
0.08
0.2
0.5
1.0


y
X
1.0
1.5
2.0
2.5
3.0
3.5
−1.0
−0.5
0.0
0.5
1.0


Mean Squared Error
1/K
0.00
0.05
0.10
0.15
0.2
0.5
1.0

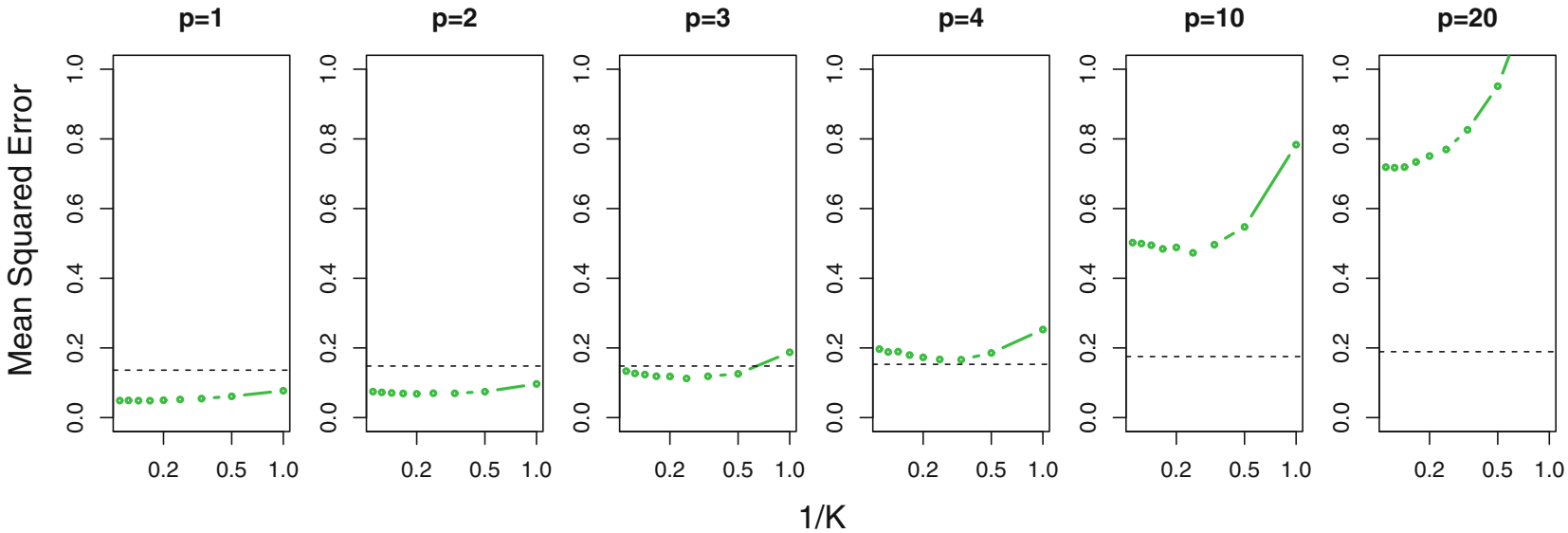
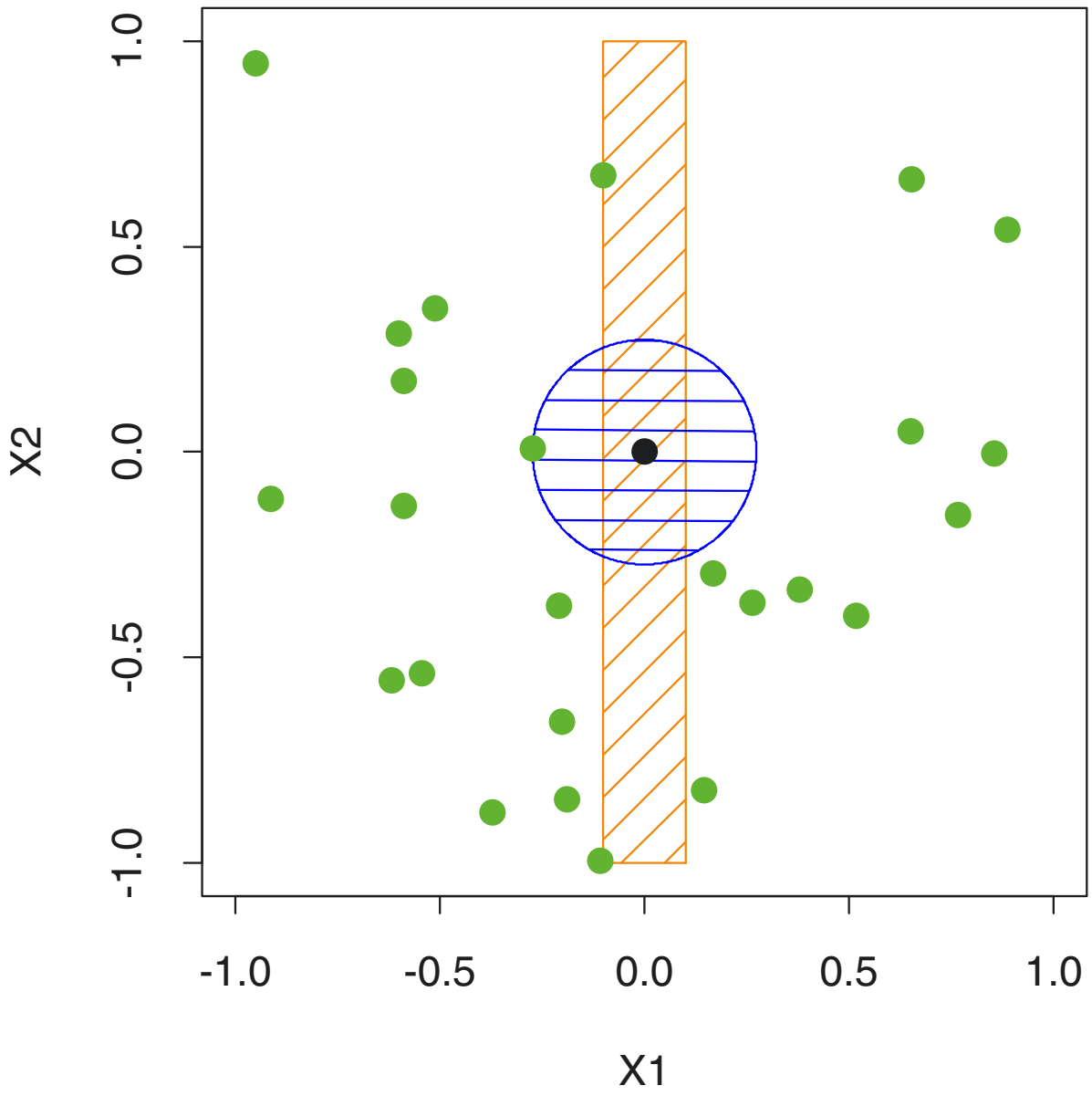


FIGURE 3.20. *Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non–linear in the first variable, as in the lower panel in Figure 3.19, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.*

Curse of Dimensionality

- When p increases, the model is more flexible, but there is fewer neighbors can be used
  - N points uniformly distributed on $[0, 1]$, average distance is $1/N$
  - N points uniformly distributed on $[0, 1]^2$, average distance is $1/\sqrt{N}$
  - To achieve the same sparsity, we need N^2 points
- Hence, as p increases linearly, N needs to grow exponentially in order to maintain the same sparsity
- Conversely, to cover the same volume, the radius needed has to increase much more faster than the dimension

1-NN in One vs. Two Dimensions



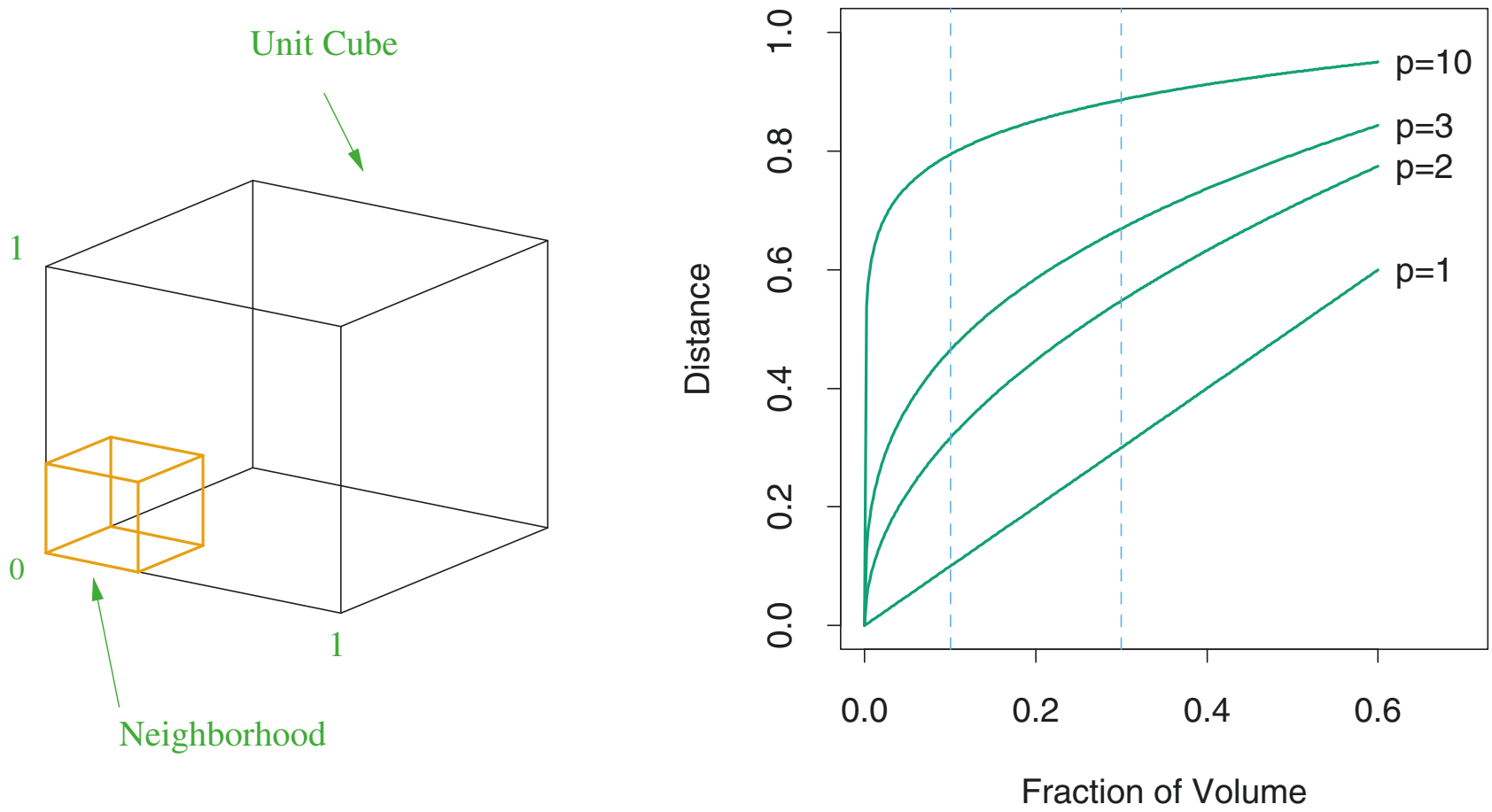


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover* 80% *of the range of each coordinate to capture* 10% *of the data.*

Classes of Restricted Estimators

- Non-parametric approach is conceptually similar to estimating infinite dimensions of parameters
- To overcome the curse of dimensionality, we need to restrict to some classes of $\hat{f}(x)$
  - Regression with Penalty: Penalize points that are too far
  - Kernal or Local Regression: Adding weights to closer points
  - Basis Expansion: Do parametric estimation locally
- In these cases, we need to determine the smoothing or complexity parameters, such as
  - The penalty term
  - Width of kernel or bands
  - The number of bases