Ch3 Regression to Machine Learning

Contents

Textbook: James et al. ISLR 2ed.

A Subsection

[ToC]

A.1 Statistical Learning

- General Model

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

- We don't want to assume that $f(X)$ is linear function.
- Motivation:
  - Model Estimation
  - Prediction
- Pattern recognition

A.2 How do we find 'overall pattern'? - Inference

- Want to understand the relationship between X and Y
- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

A.3 Prediction

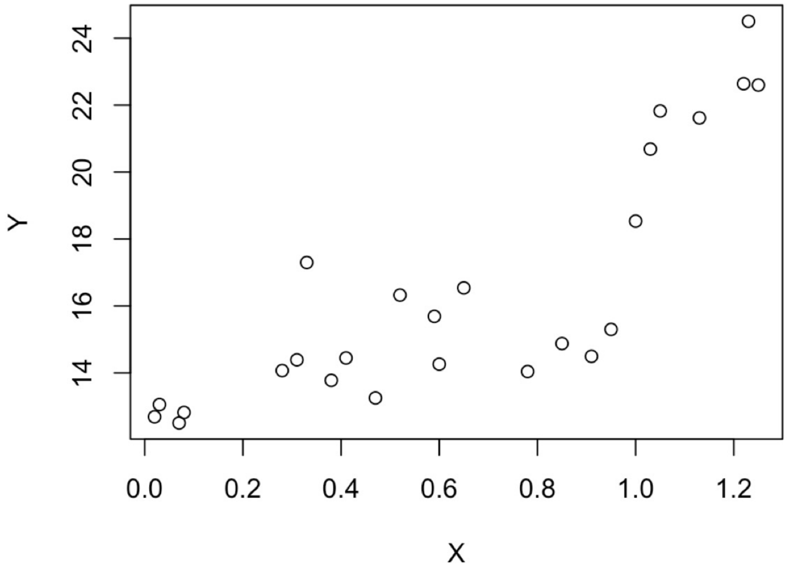
- Want to guess the next Y as accurate as possible

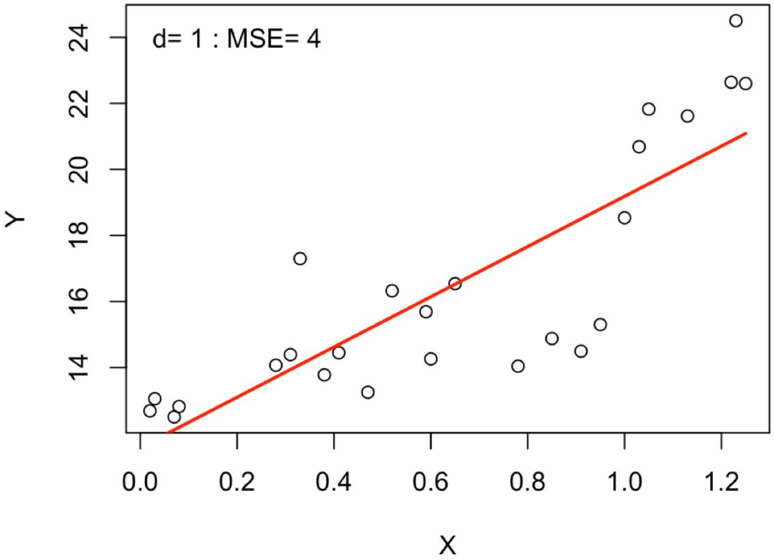
$$\hat{Y} = \hat{f}(X)$$

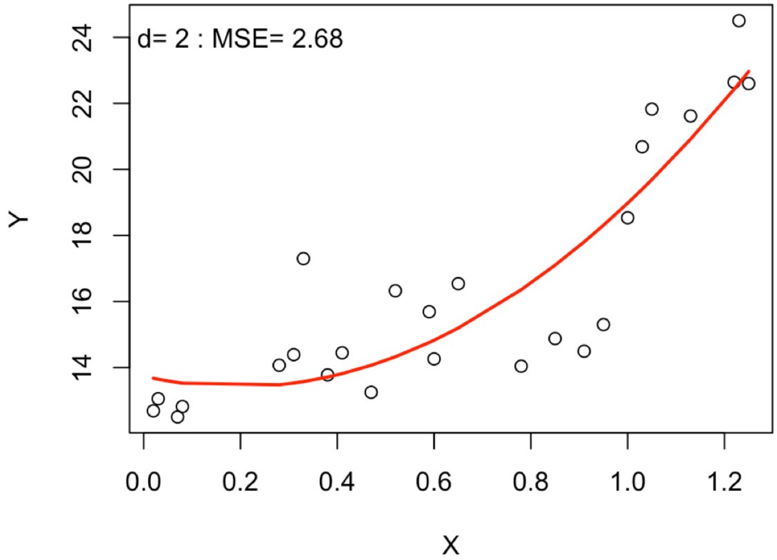
- f can be a black box
- reducible error and irreducible error in prediction
- Want to reduce prediction Mean Squared Error:

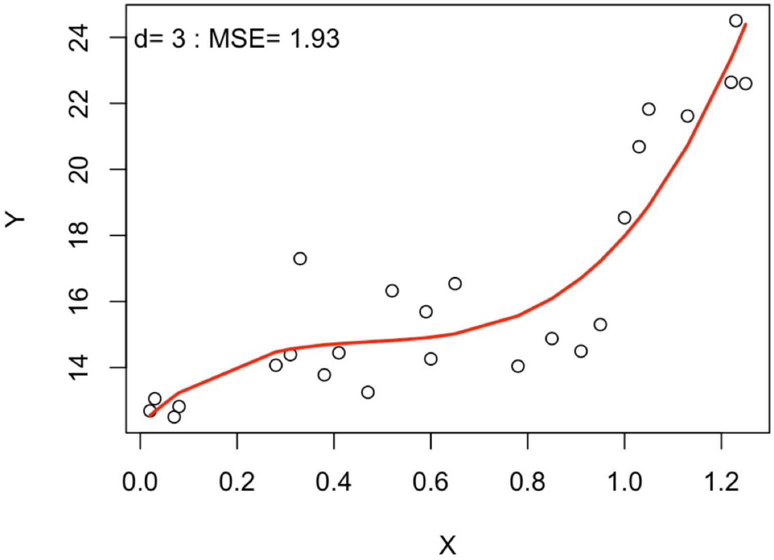
$$MSE = E(Y - \hat{Y})^2 = E(Y - \hat{f}(X))^2$$

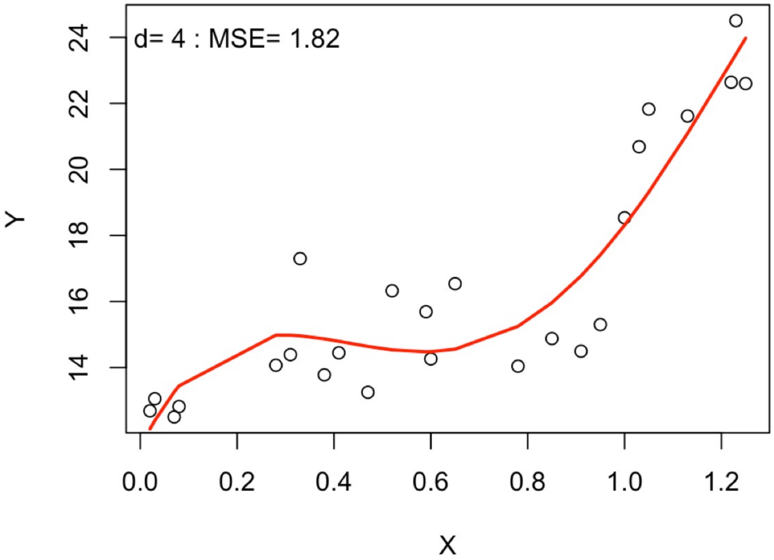
A.4 Polynomial Regression 1

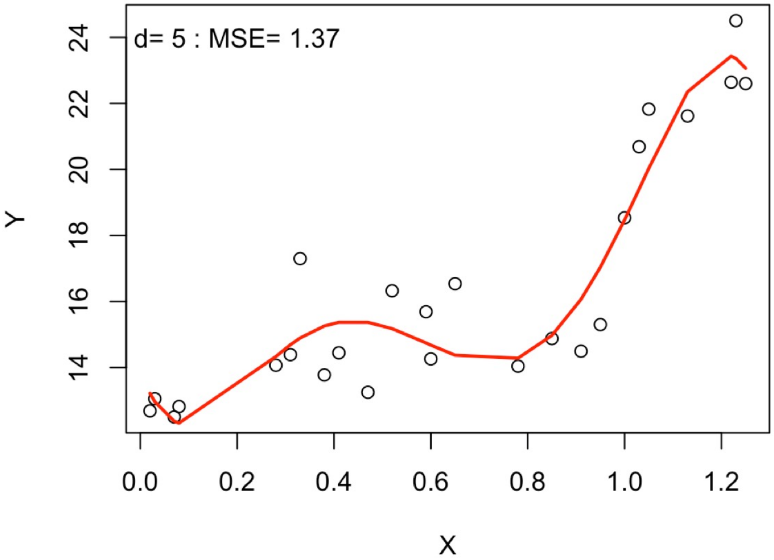


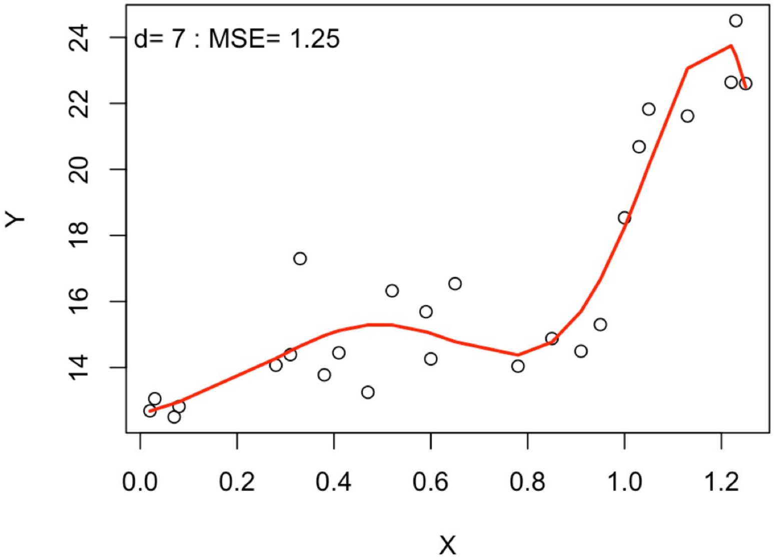

d= 1 : MSE= 4
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X


d= 2 : MSE= 2.68
Y
14
16
18
20
22
24
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2

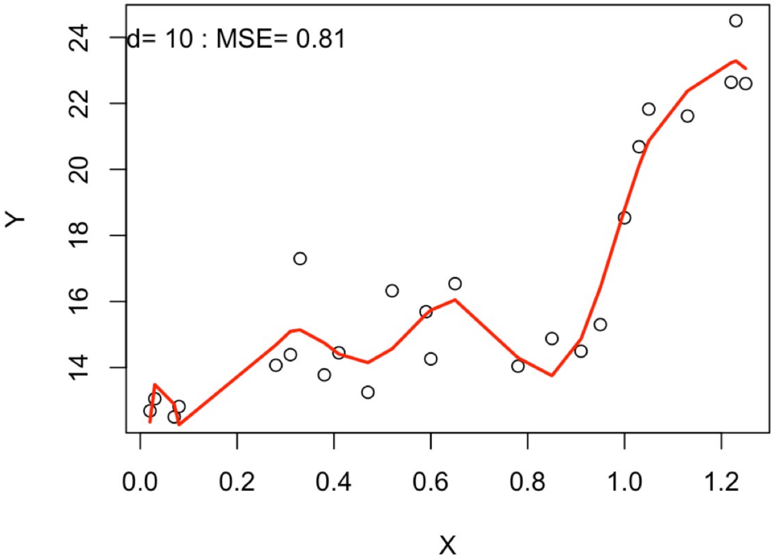

d= 3 : MSE= 1.93
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X

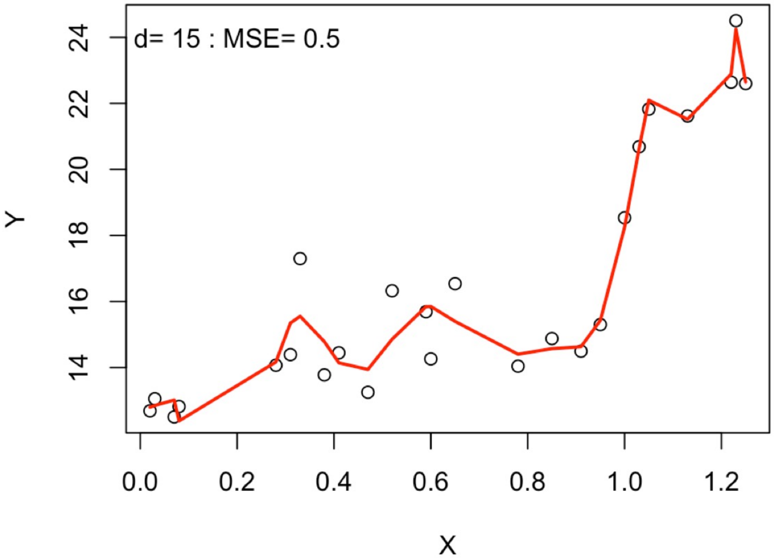

d= 4 : MSE= 1.82
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X

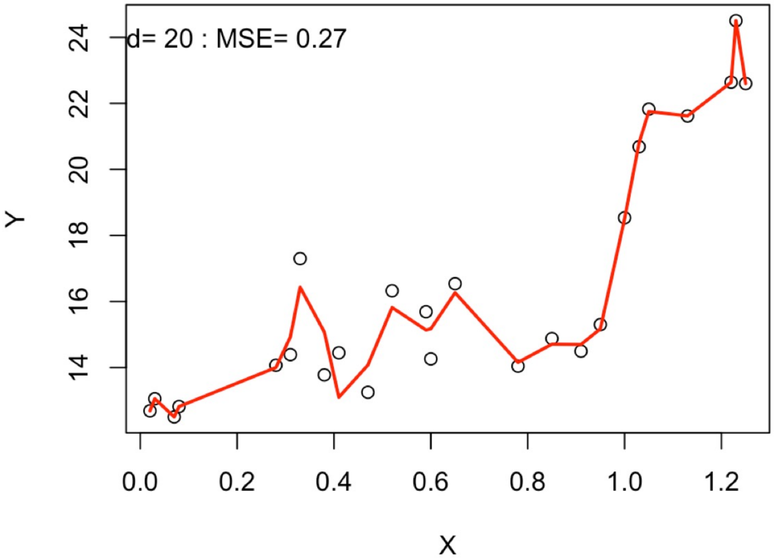

d= 5 : MSE= 1.37
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X



d= 7 : MSE= 1.25
Y
X
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2

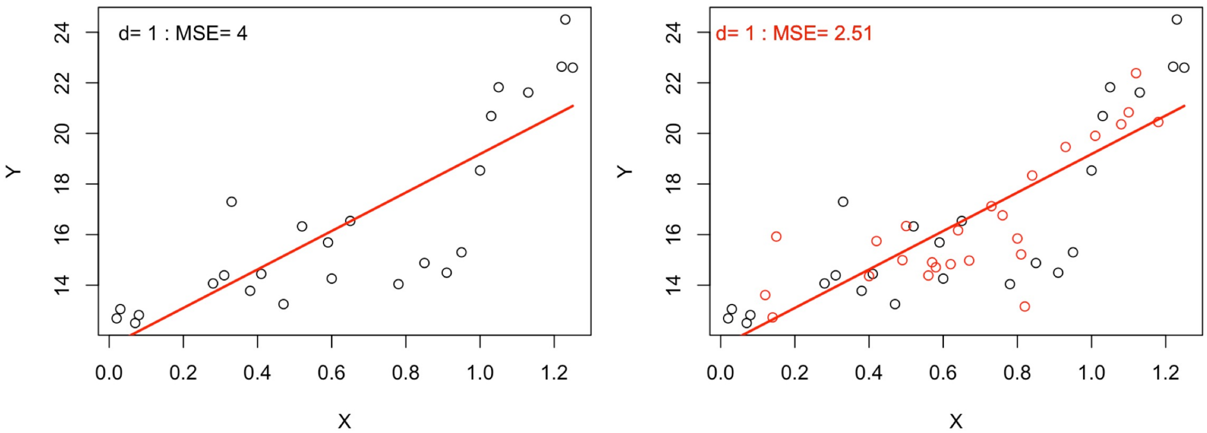

d= 10 : MSE= 0.81
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X

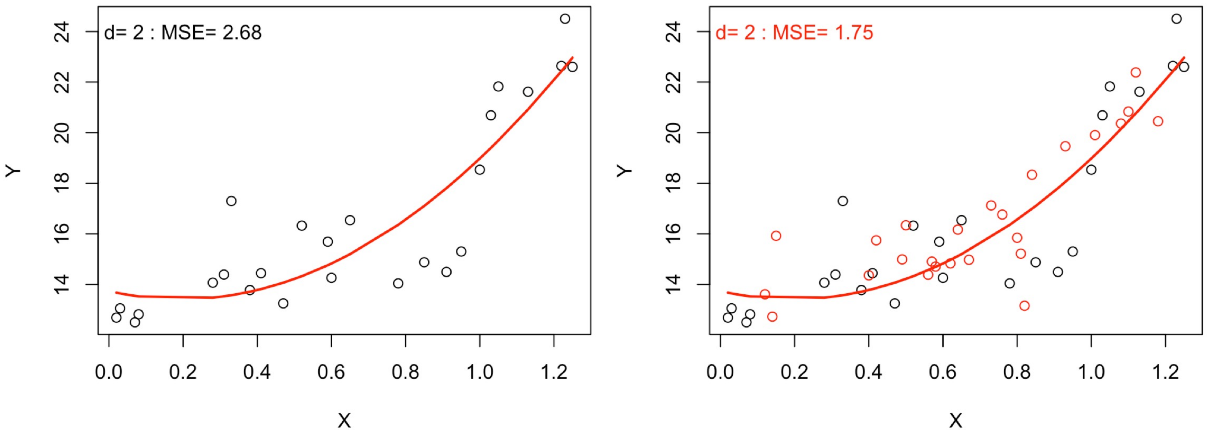

d= 15 : MSE= 0.5
Y
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X



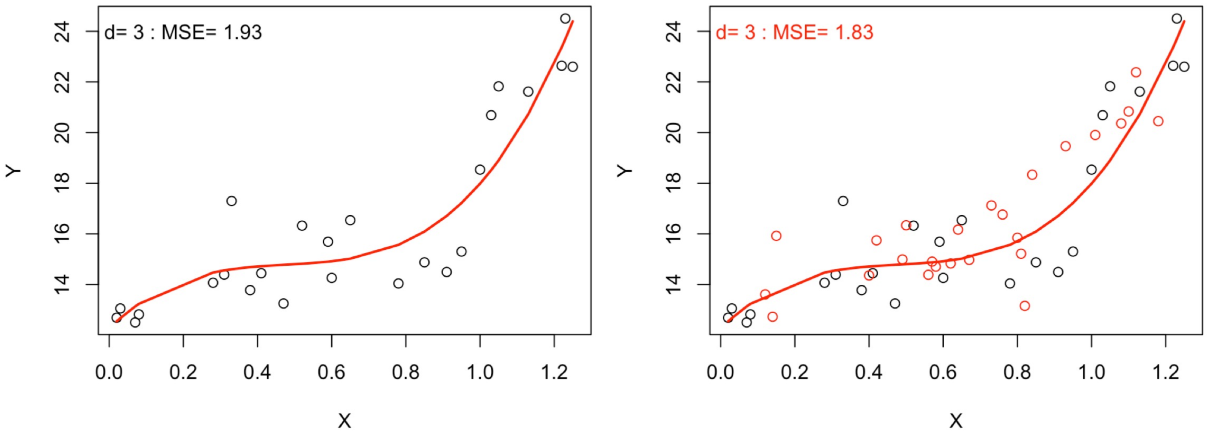
d= 20 : MSE= 0.27
Y
X

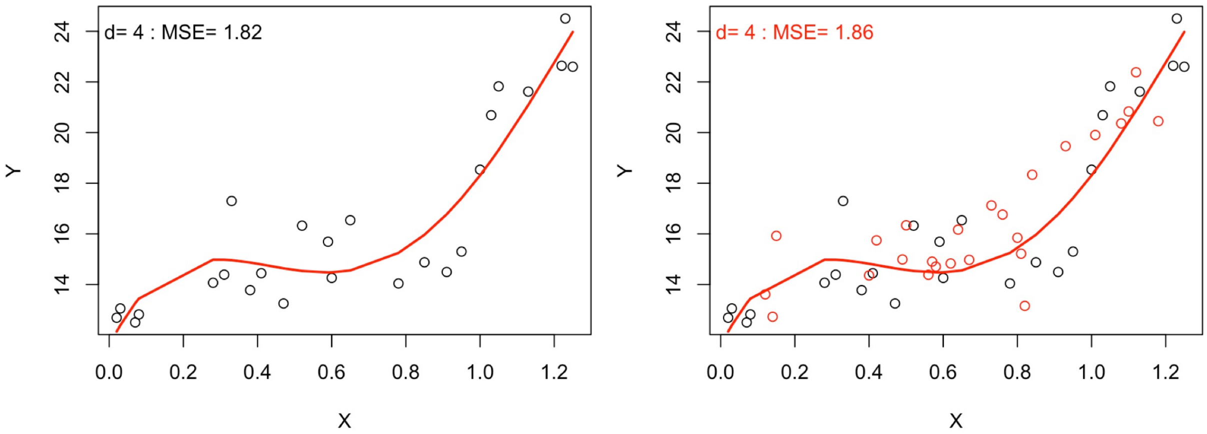
A.5 Prediction



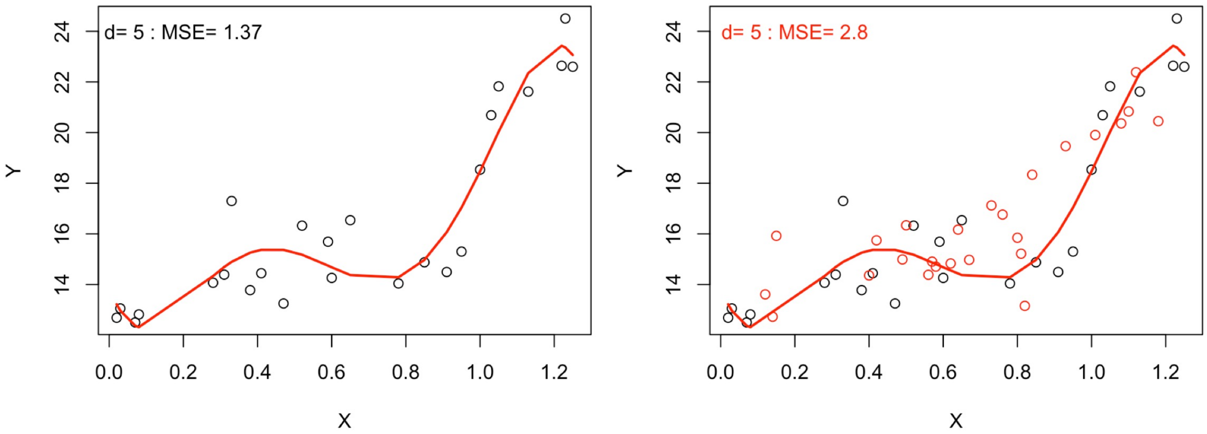


d= 2 : MSE= 2.68
Y
X
d= 2 : MSE= 1.75
Y
X

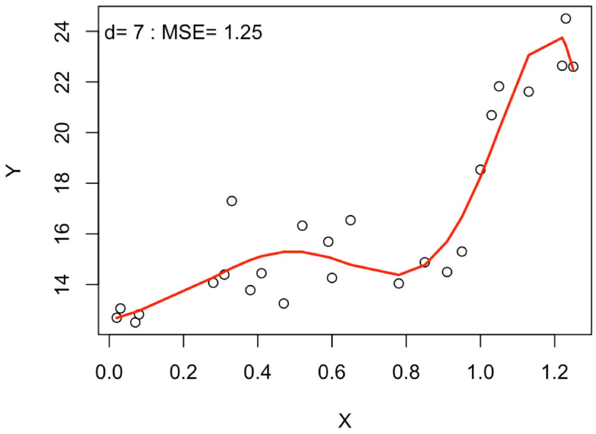

d= 3 : MSE= 1.93
d= 3 : MSE= 1.83
Y
X



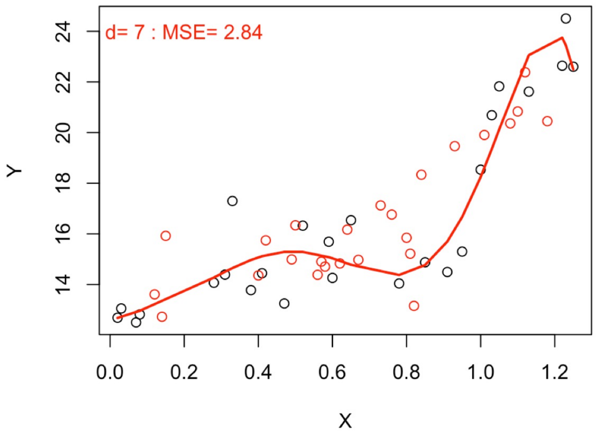
d= 4 : MSE= 1.82
d= 4 : MSE= 1.86
Y
X
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2



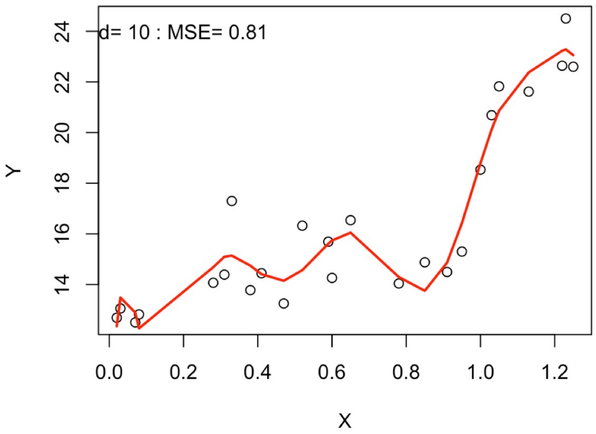
d= 5 : MSE= 1.37
d= 5 : MSE= 2.8
Y
X



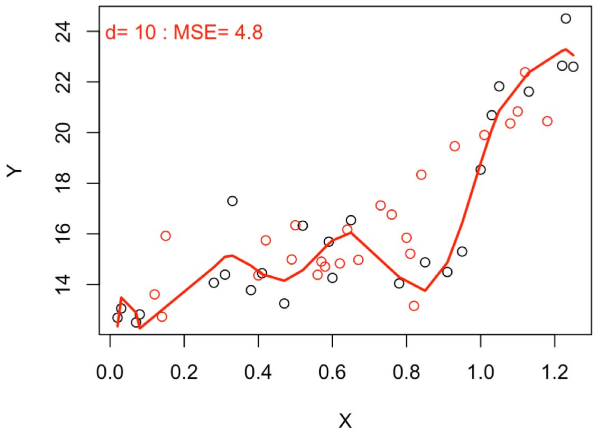
d= 7 : MSE= 1.25
Y
X



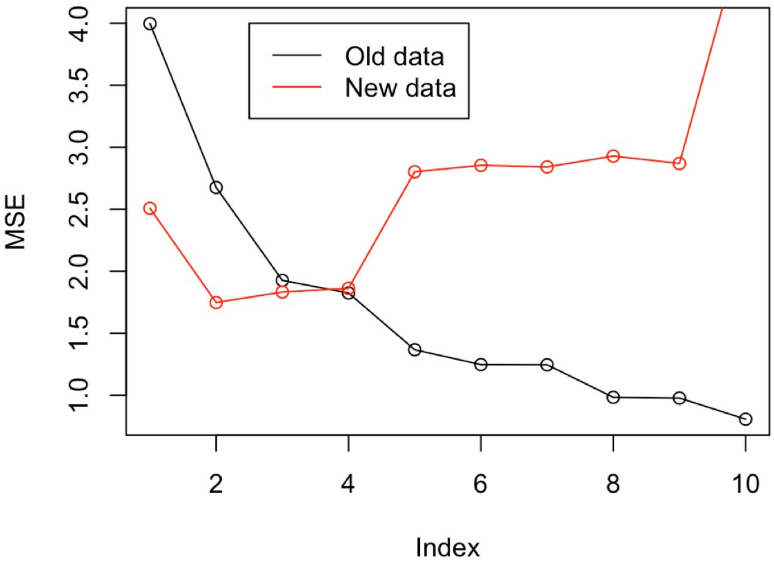
d= 7 : MSE= 2.84
Y
X



d= 10 : MSE= 0.81
Y
X

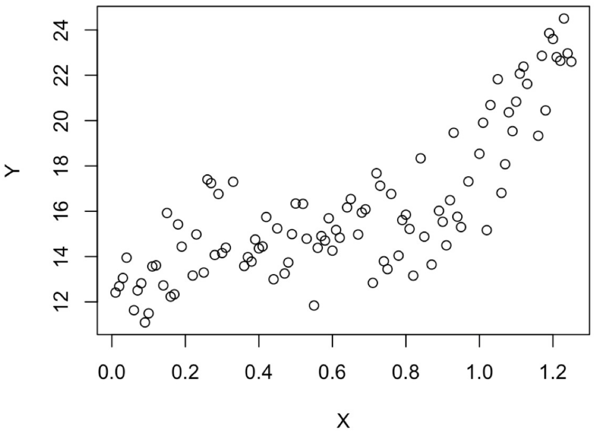


d= 10 : MSE= 4.8
Y
X

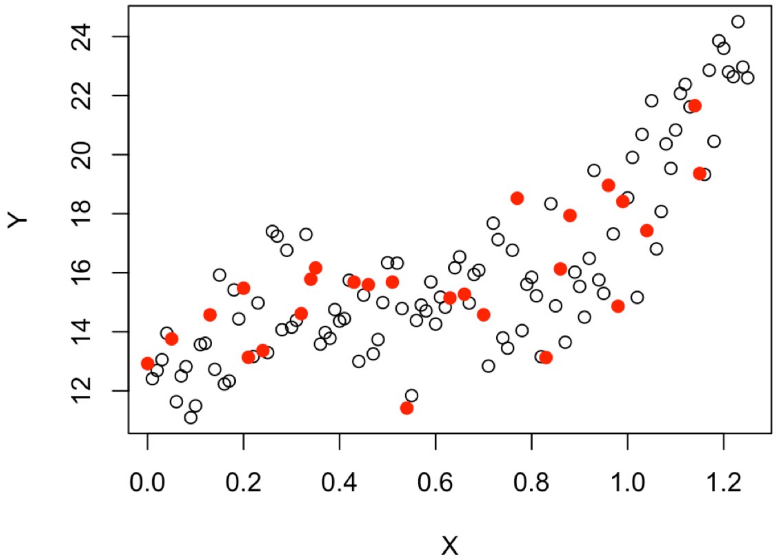

Old data
New data
MSE
Index
1.0
1.5
2.0
2.5
3.0
3.5
4.0
2
4
6
8
10

A.6 Polynomial Regression 2

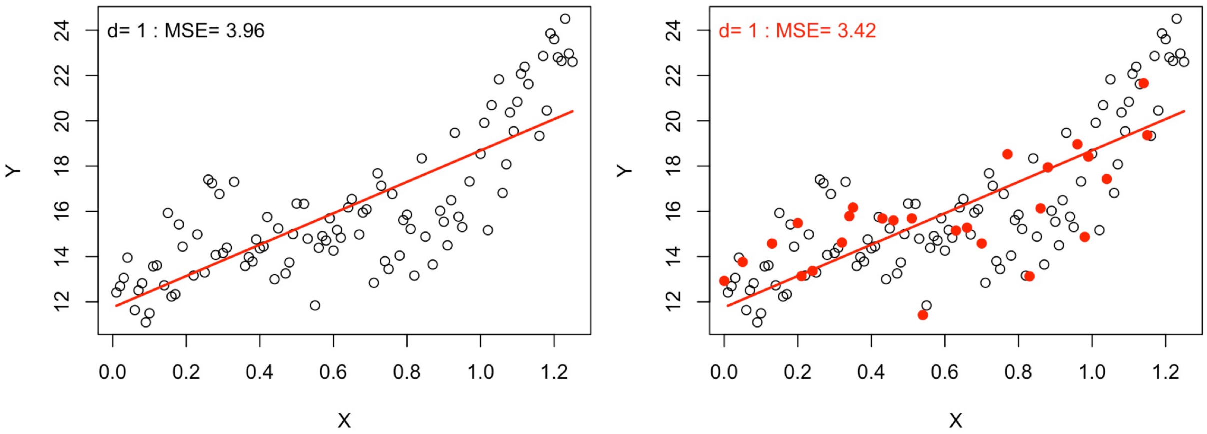
Larger dataset. $n = 100$ and $m = 26$.



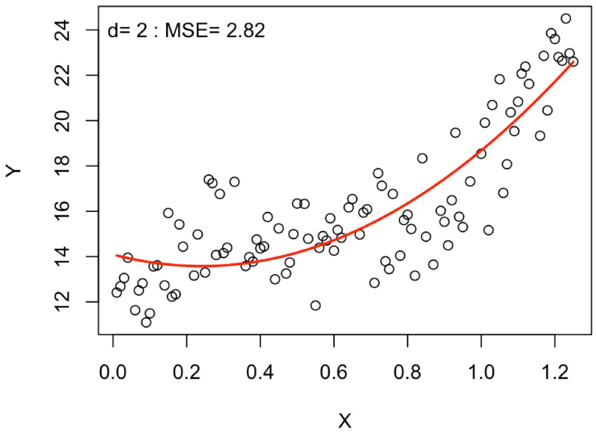
Y
X
12
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2

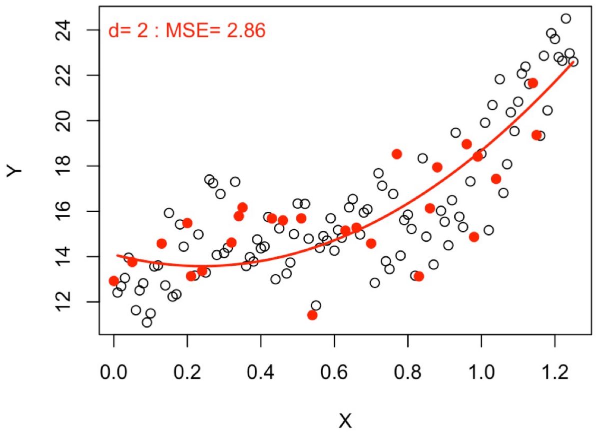


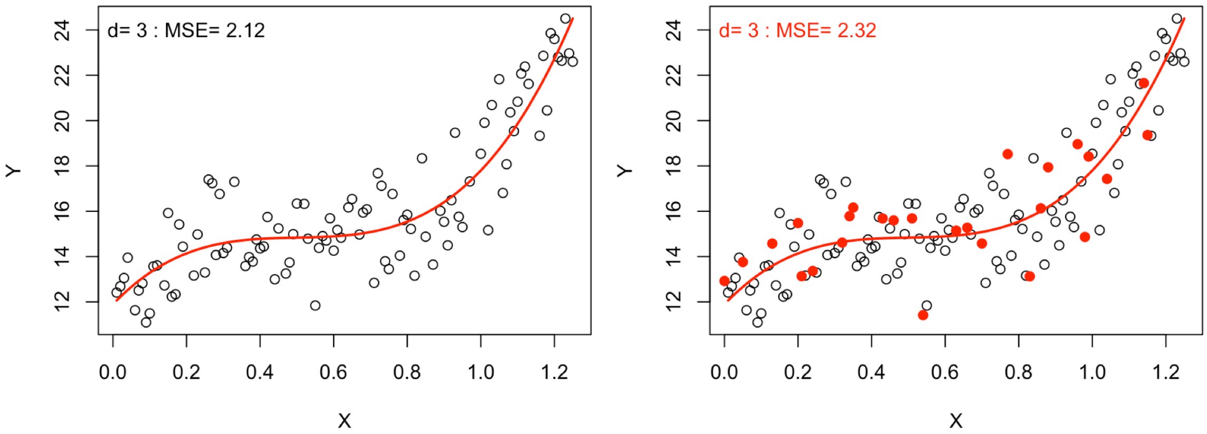
Y
X
12
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2

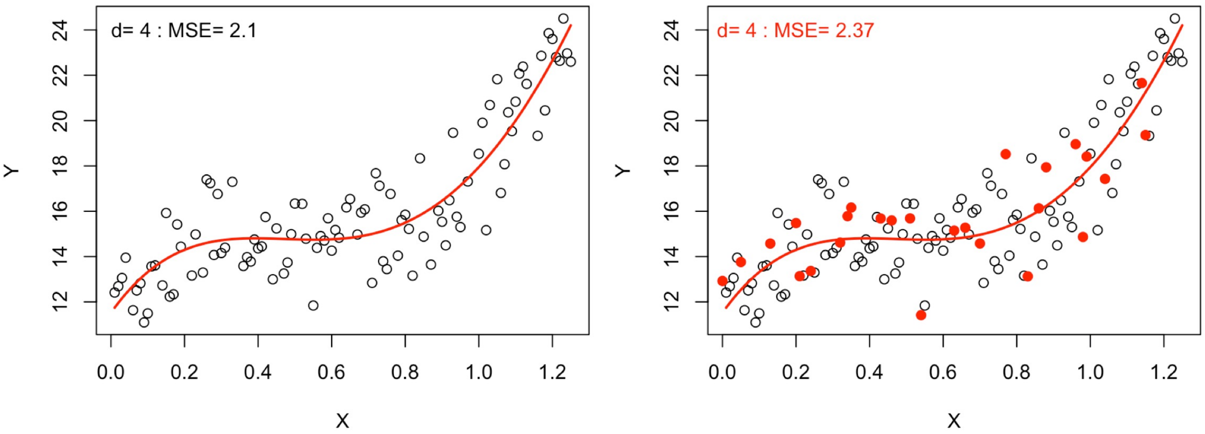


d= 1 : MSE= 3.96
d= 1 : MSE= 3.42
Y
X
24
22
20
18
16
14
12
0.0
0.2
0.4
0.6
0.8
1.0
1.2

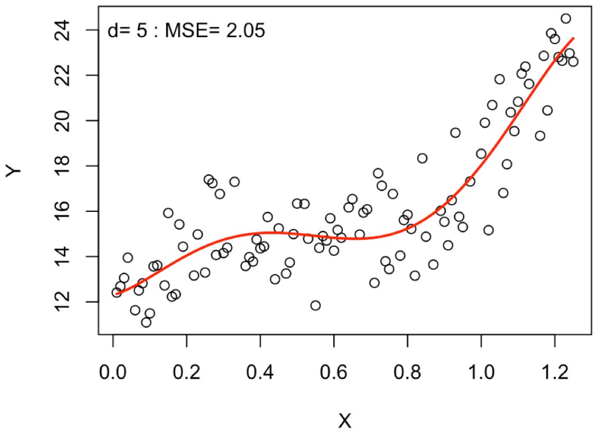

d= 2 : MSE= 2.82
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
12
14
16
18
20
22
24

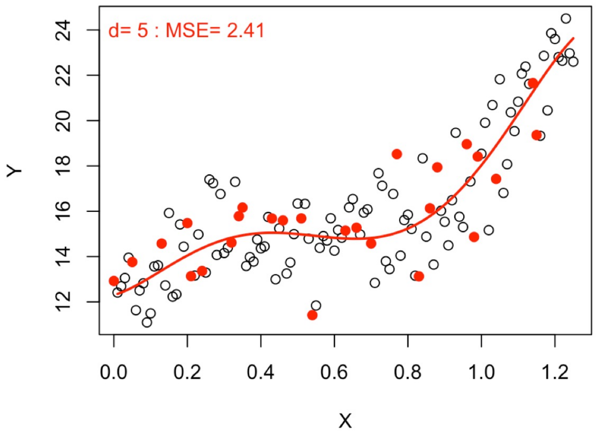

d= 2 : MSE= 2.86
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
12
14
16
18
20
22
24

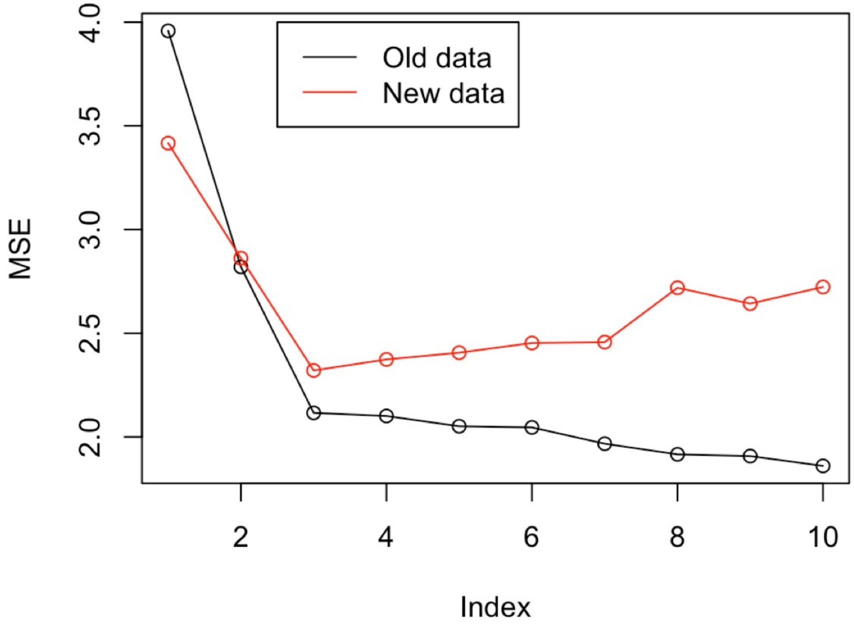

d= 3 : MSE= 2.12
d= 3 : MSE= 2.32
Y
X
12
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2



d= 4 : MSE= 2.1
d= 4 : MSE= 2.37
Y
X


d= 5 : MSE= 2.05
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
12
14
16
18
20
22
24


d= 5 : MSE= 2.41
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
12
14
16
18
20
22
24



Old data
New data
MSE
Index
4.0
3.5
3.0
2.5
2.0
2
4
6
8
10

A.7 Training, Validation, and Testing Set

- k-fold Cross Validation

A.8 How do we estimate f?

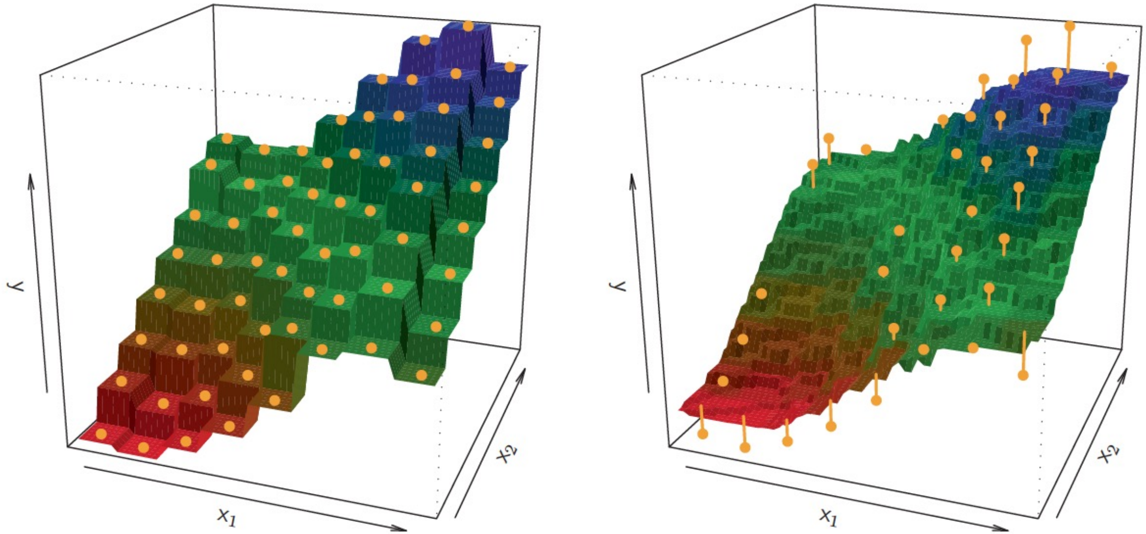
- Hyperparameter

A.9 K-Nearest Neighbor

- Pick a point x_0
- Find K nearest observations
- $f(x_0)$ is estimated by the average of all K neighbors

$$\hat{f}(x_0) = \frac{1}{K} \sum y_i$$

- K=1 (left) and K=9 (right)



y
x_1
x_2
y
x_1
x_2

A.10 Assessing Model Prediction Accuracy

Assessing Model Prediction Accuracy

A.11 Measure of Quality of Fit

- Training MSE (sample)

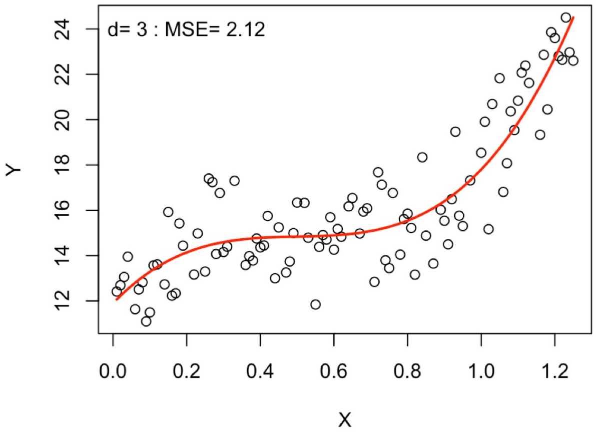
$$\text{MSE}_{tr} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

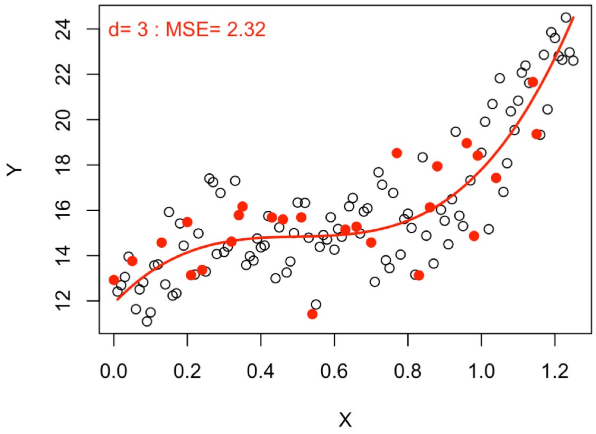
- But we want minimum Prediction MSE

$$\text{MSE} = E\left(Y - \hat{f}(X)\right)^2$$

- Solution: look at Test MSE (sample) as estimator

$$\text{MSE}_{test} = \frac{1}{m} \sum_{j=1}^{m} \left(y_j - \hat{f}(x_j) \right)^2$$





A.12 Bias-Variance Trade-Off

Bias-Variance Trade-Off

A.13 Prediction MSE

$$E\big(Y - \hat{f}(X)\big)^2 = Var(\hat{f}(X)) + Bias(\hat{f}(X))^2 + Var(\epsilon)$$

- can't have low variance and low bias
- has lower bound

Bias-Variance Trade-off

Prediction MSE can be decomposed as

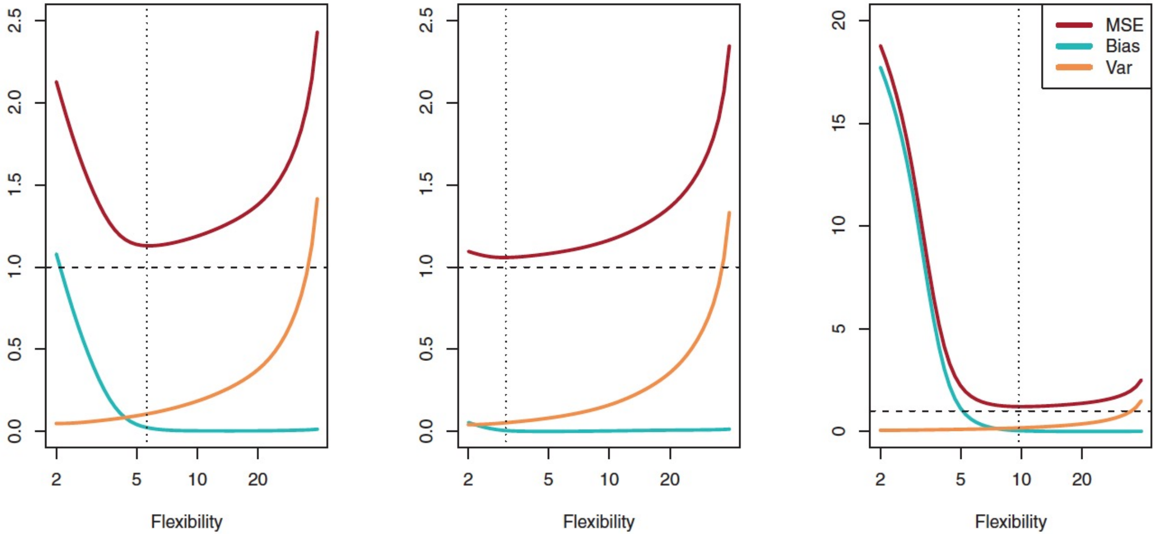
$$E\Big(Y - \hat{f}(X)\Big)^2 = E\bigg(f(X) + \epsilon - \hat{f}(X)\bigg)^2$$

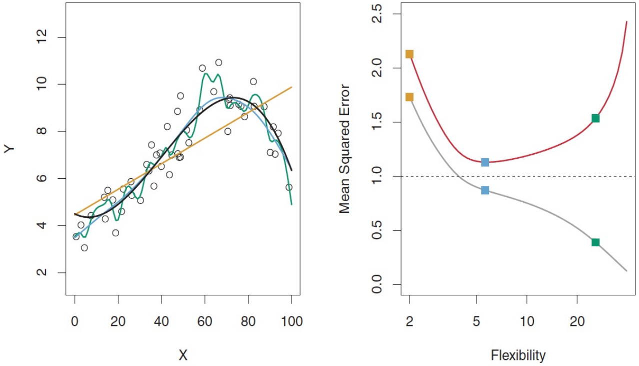
$$= E\bigg(f(X) - E(\hat{f}(X)) + E(\hat{f}(X)) - \hat{f}(X) + \epsilon\bigg)^2$$

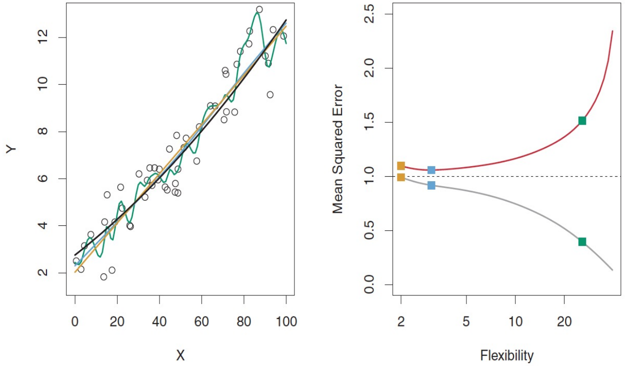
$$= E\bigg(f(X) - E(\hat{f}(X))\bigg)^2 + E\bigg(E(\hat{f}(X)) - \hat{f}(X)\bigg)^2 + E\Big(\epsilon^2\Big)$$

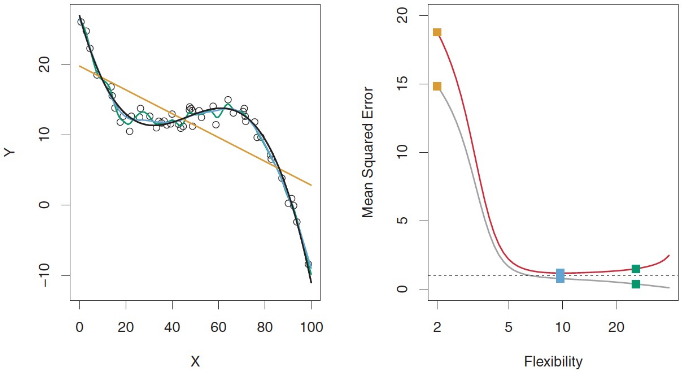
$$= Var(\hat{f}(X)) \; + \; Bias(\hat{f}(X))^2 \; + \; Var(\epsilon)$$

A.14 Regression vs Classification




Y
X
Mean Squared Error
Flexibility

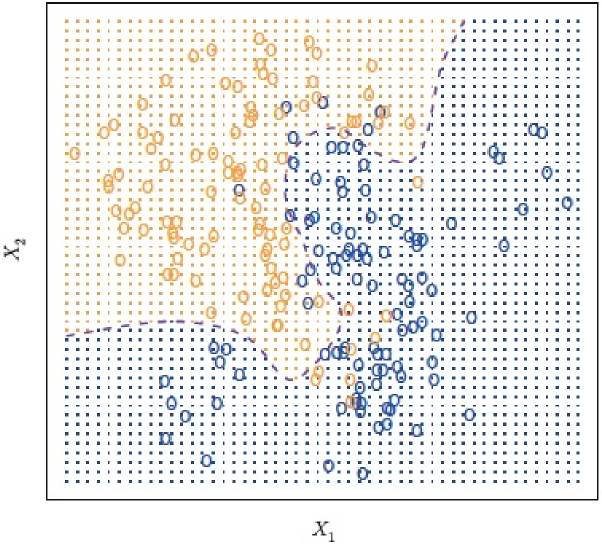

Y
X
Mean Squared Error
Flexibility


Y
X
Mean Squared Error
Flexibility

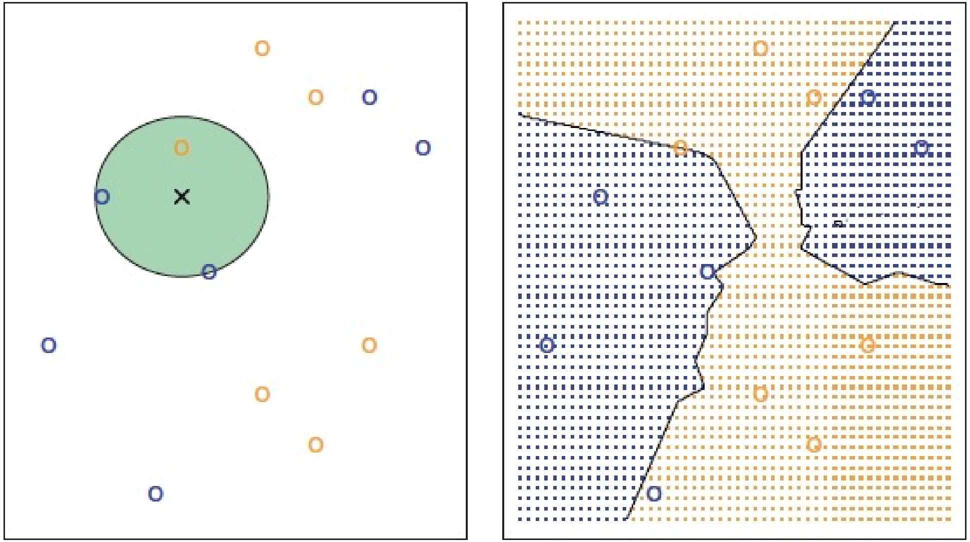
A.15 Classification Setting

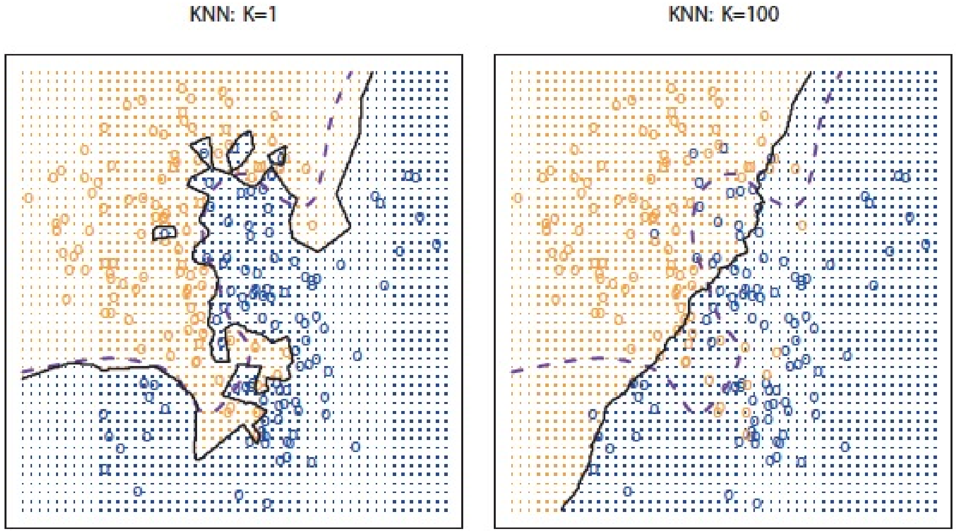
- Instead of MSE, work with Errror Rate:

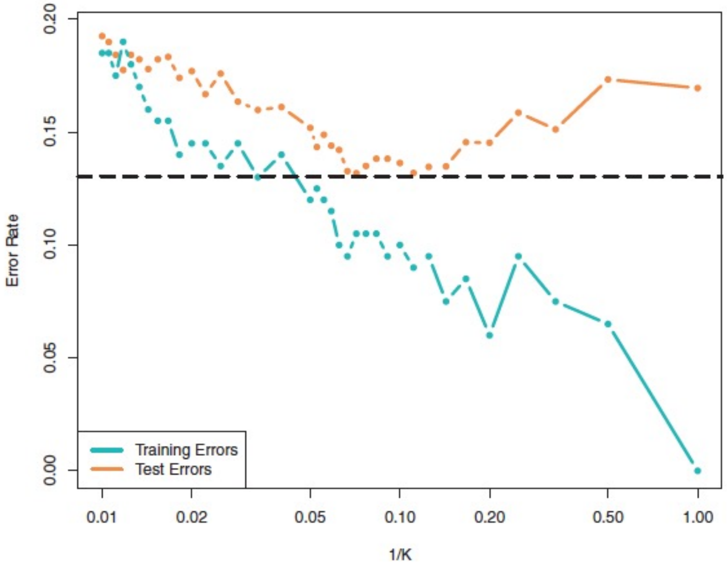
$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$



X_2
X_1




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