Ch3 Regression to Machine Learning

Contents

Textbook: James et al. ISLR 2ed.

3 Subsection

[ToC]

A.1 Statistical Learning

- General Model

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

- We don't want to assume that $f(X)$ is linear function.
- Two types of 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 How do we find 'overall pattern'? - Prediction

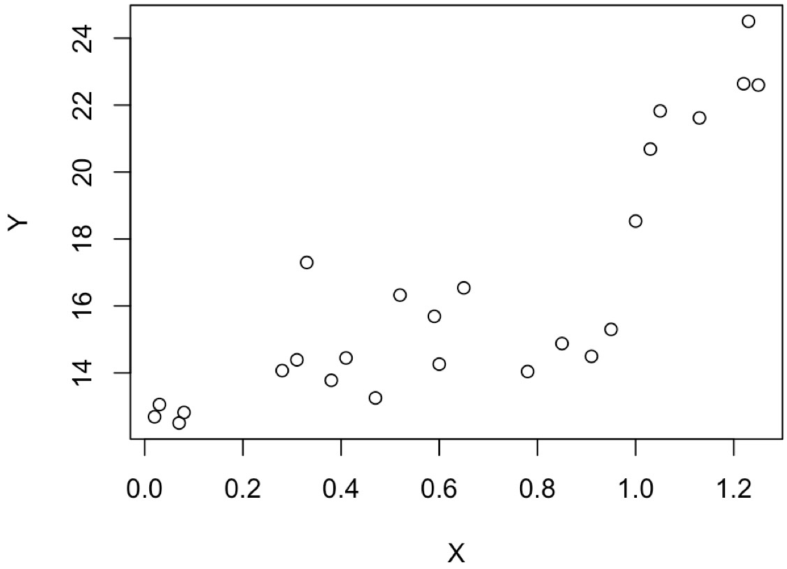
- Want to guess the next Y as accurate as possible

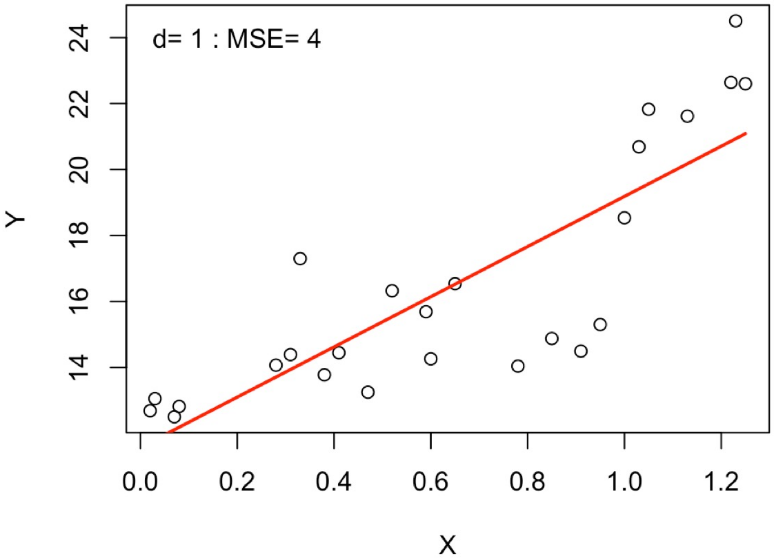
$$\hat{Y} = \hat{f}(X)$$

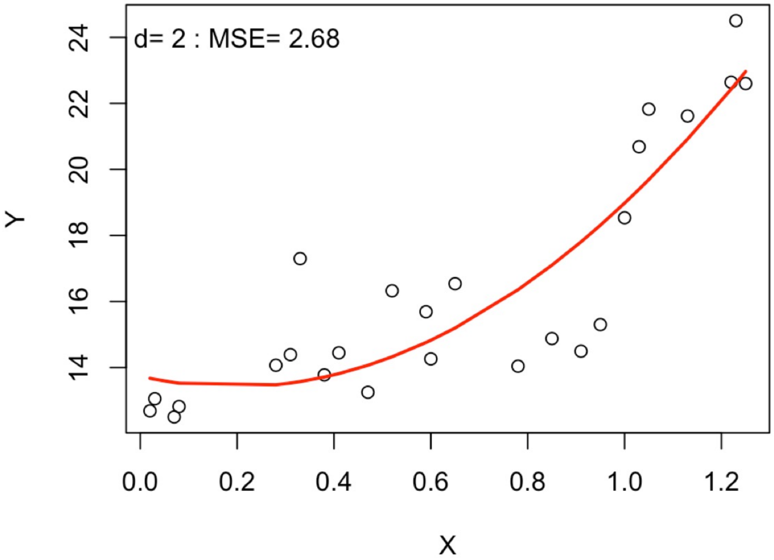
- f can be a black box
- reducible error and irreducible error in prediction
- Want to reduce prediction Mean Squared Error:

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

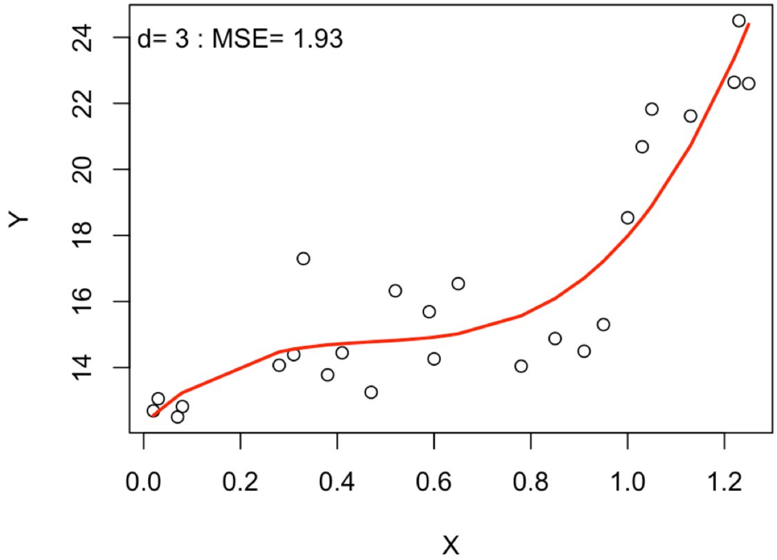
A.4 Polynomial Regression 1

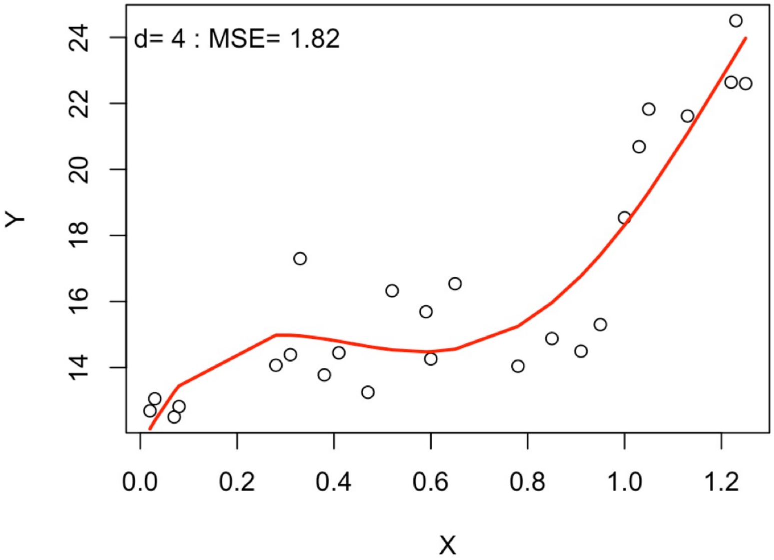


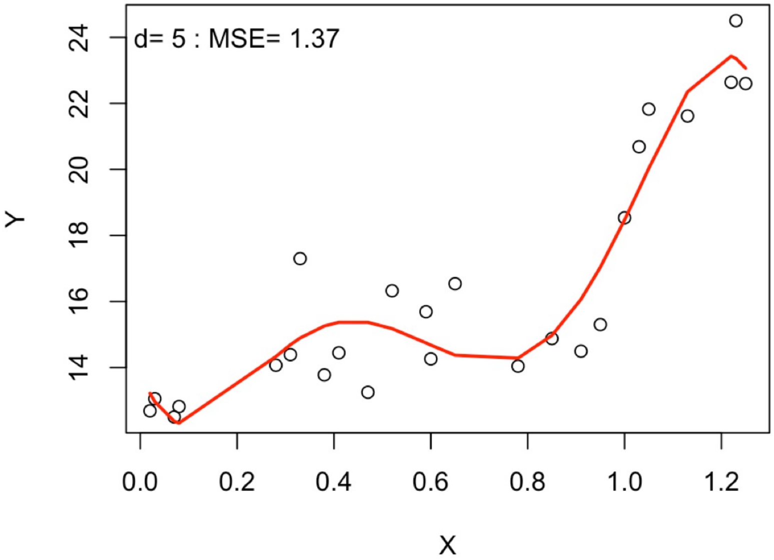

d= 1 : MSE= 4
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X



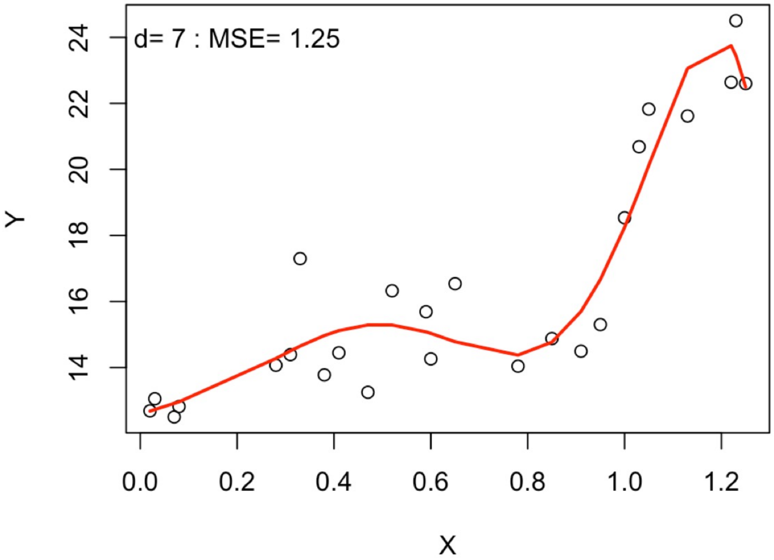
d= 2 : MSE= 2.68
Y
14
16
18
20
22
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X

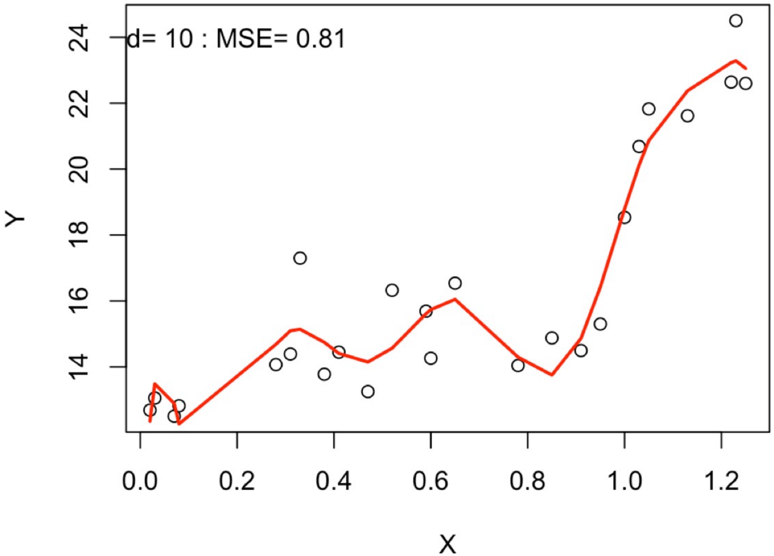

d= 3 : MSE= 1.93
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X

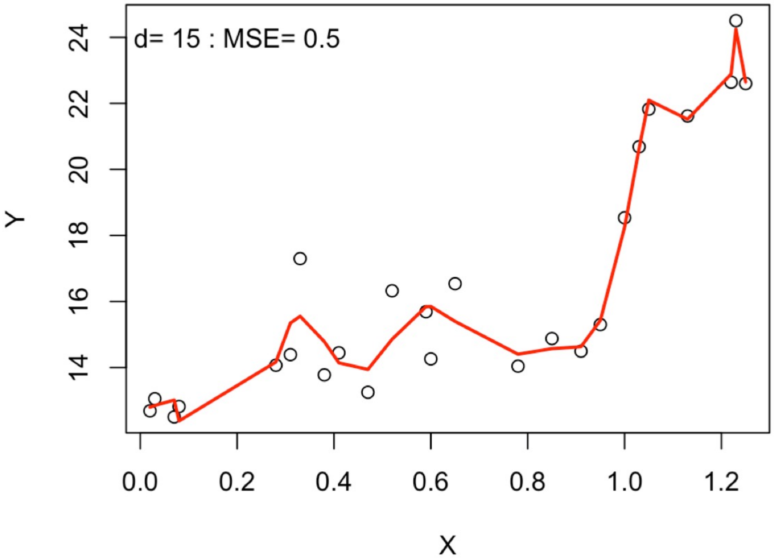

d= 4 : MSE= 1.82
Y
X
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2



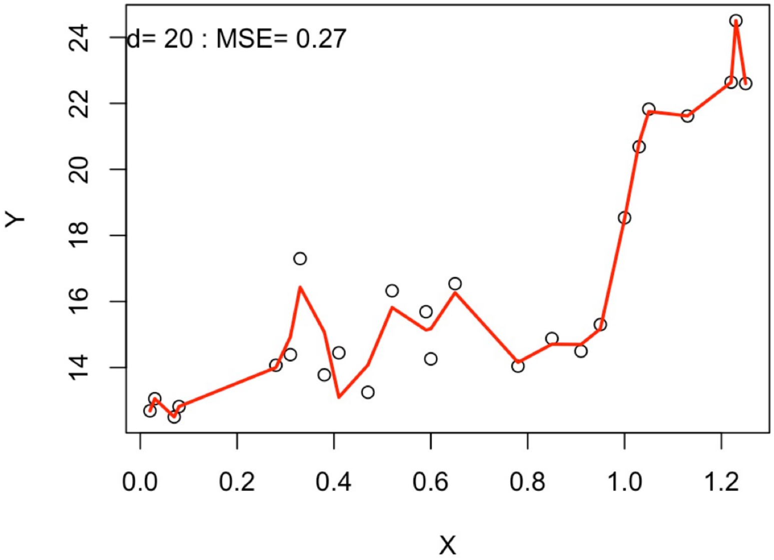
d= 5 : MSE= 1.37
Y
24
22
20
18
16
14
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X


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


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



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



d= 20 : MSE= 0.27
Y
14
16
18
20
22
24
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2

A.5 Problem

- More flexibility in the model is always going to result in better fit to the data.
- Better fitting model is not always inferential.
- Better fitting Leave some out and use it for 'validation' and 'testing'.
- Underlying mechanism:

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

A.6 Measure of Quality of Fit

- Training MSE (sample)

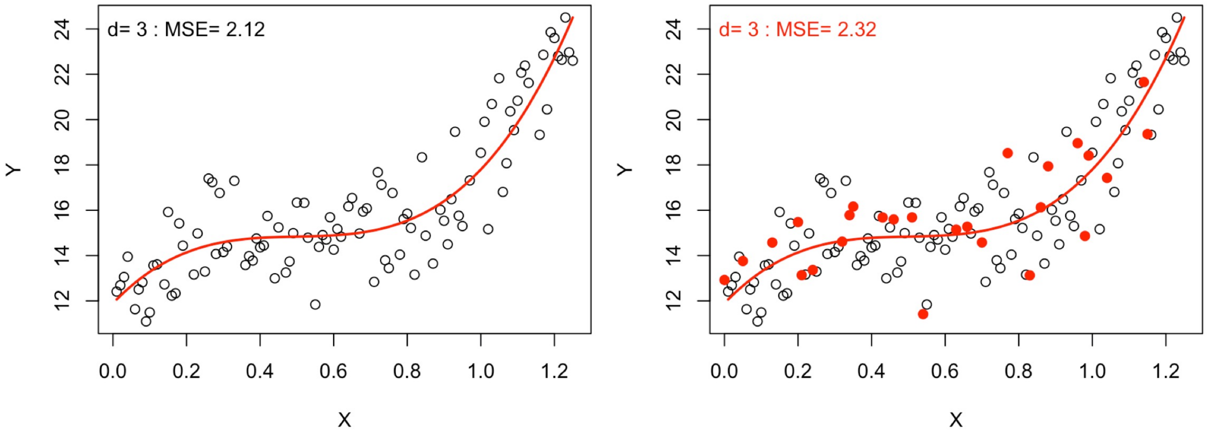
$$\text{MSE}_{tr} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

- But we want minimum Prediction MSE

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

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

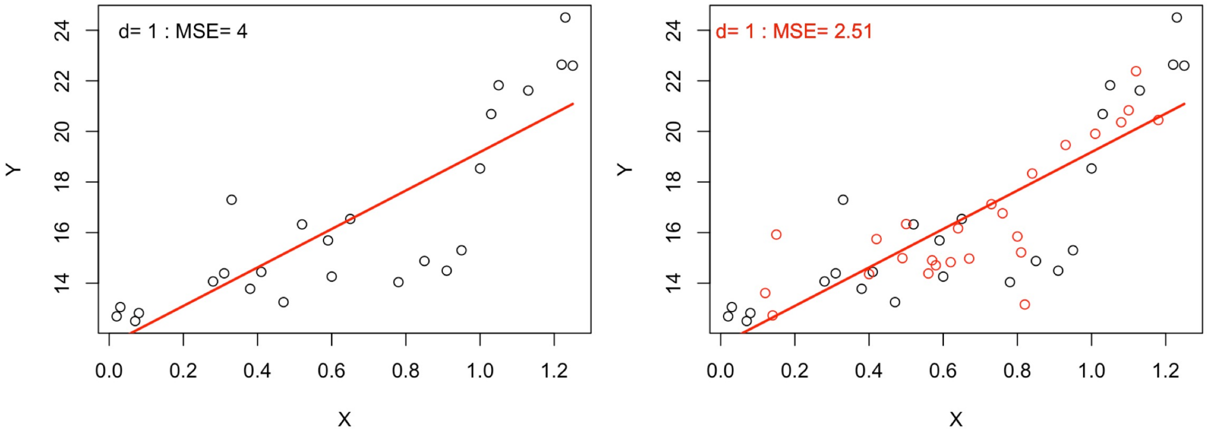
$$\text{MSE}_{test} = \frac{1}{m}\sum_{j=1}^{m}\left(y_j - \hat{f}(x_j)\right)^2$$

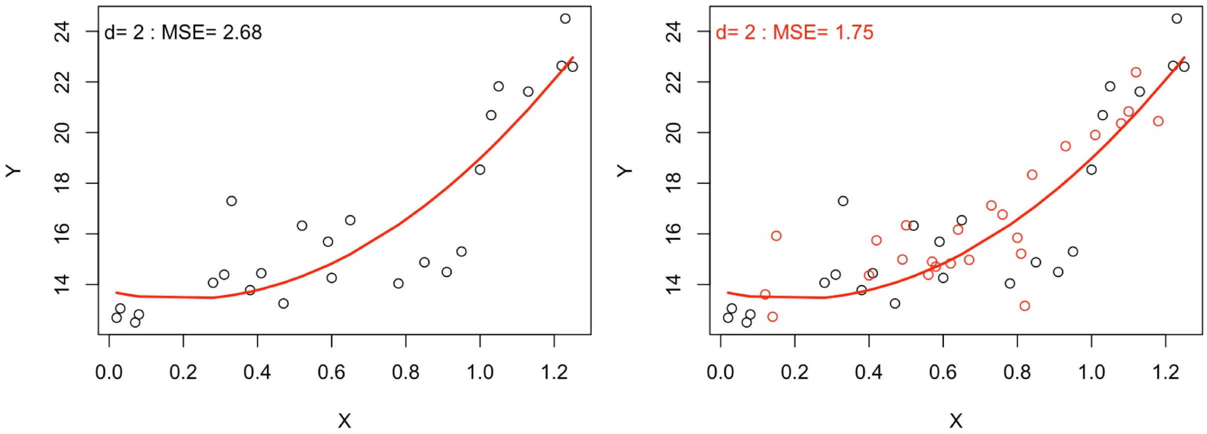


A.7 KEY CONCEPT

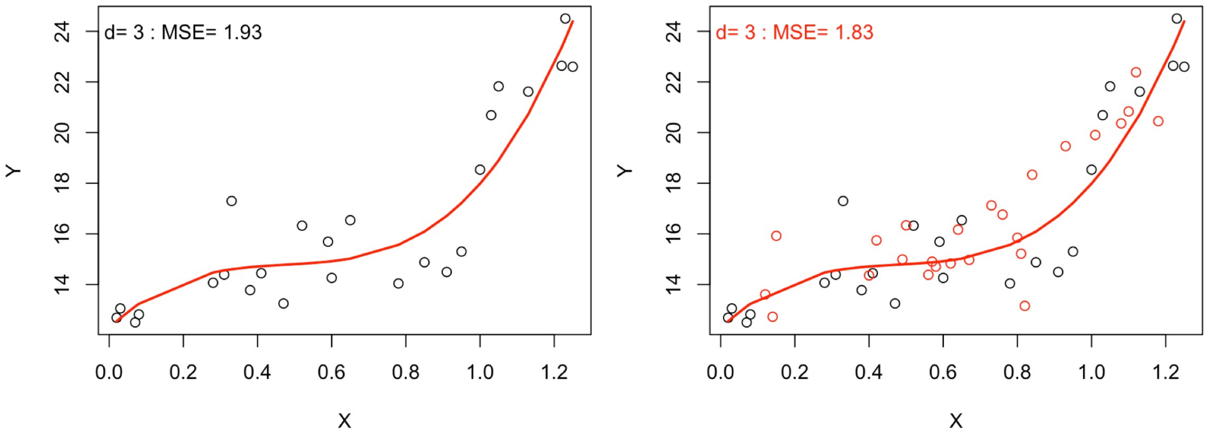
- Cross Validation
- Don't use all data when you are fitting a model
- Leave some out and use it for 'validation' and 'testing'.

A.8 Leave-some-out Fitting Procedure 1

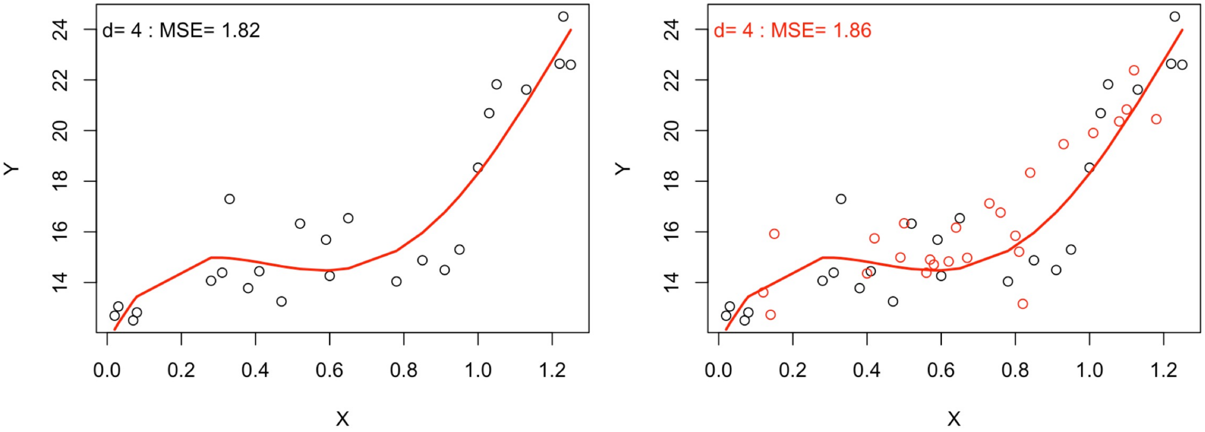


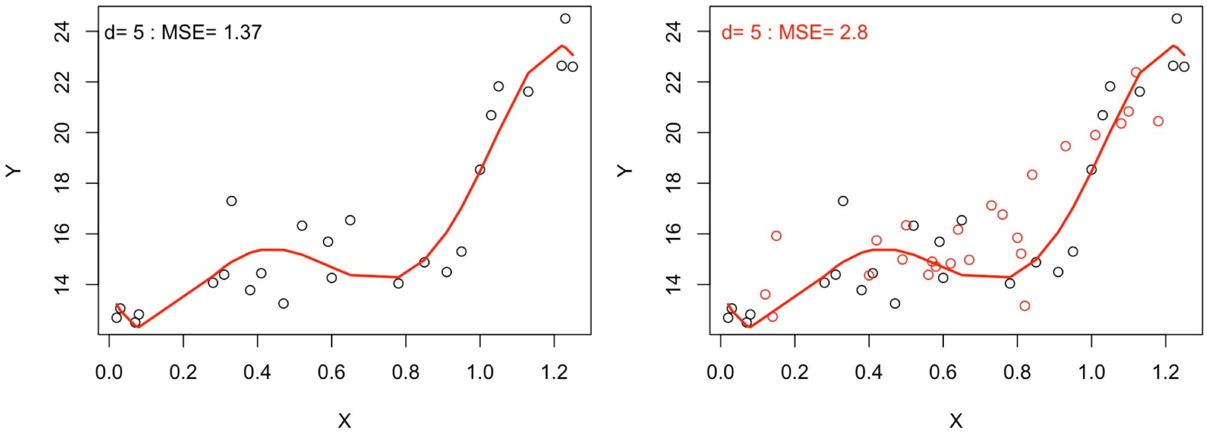


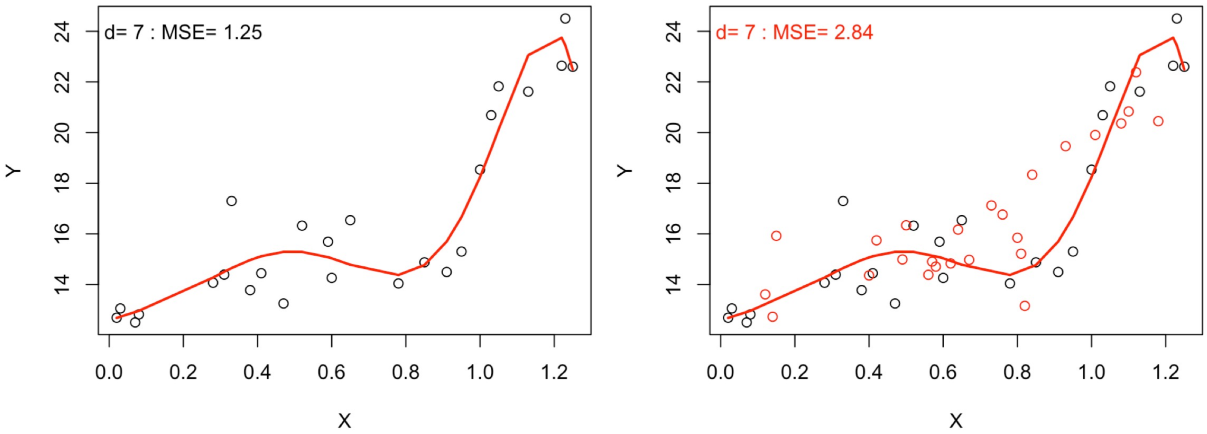
d= 2 : MSE= 2.68
d= 2 : MSE= 1.75
Y
X



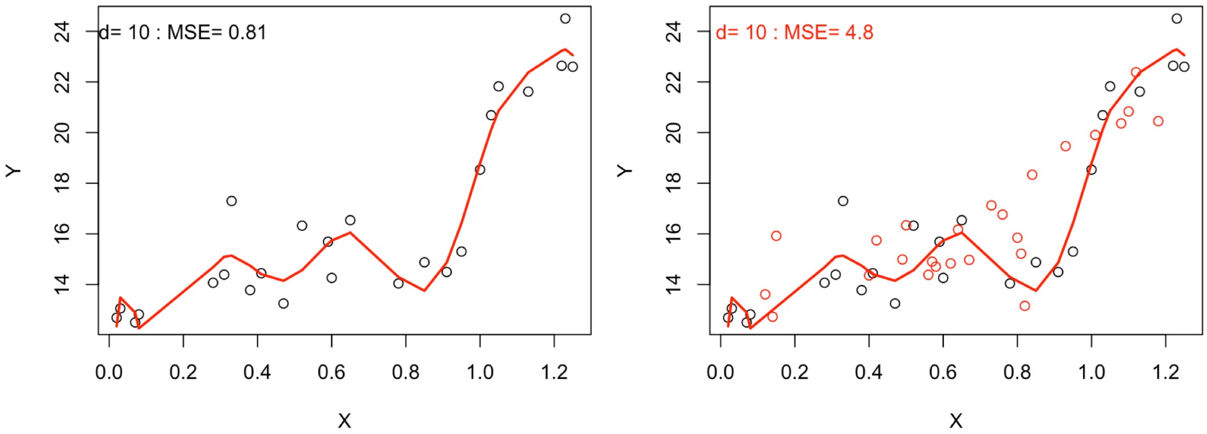
d= 3 : MSE= 1.93
d= 3 : MSE= 1.83
Y
X


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

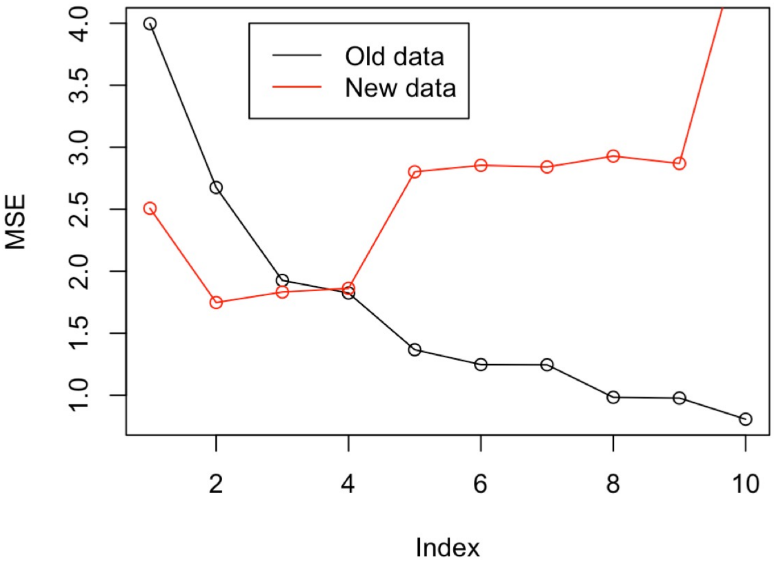

d= 5 : MSE= 1.37
d= 5 : MSE= 2.8
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
1.2
14
16
18
20
22
24



d= 7 : MSE= 1.25
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
4.0
3.5
3.0
2.5
2.0
1.5
1.0
2
4
6
8
10

A.9 KEY CONCEPT

- Divide data into

$$
\begin{aligned}
[\text{Training}] &\quad \text{vs} \quad [\text{Testing}] \\
[\text{Training}] &\quad \text{vs} \quad [\text{Validation}] \\
[\text{In-sample}] &\quad \text{vs} \quad [\text{Out-sample}]
\end{aligned}
$$

- Fit the model using [Training] data
- See if the model actually fit [Testing] data (the model hasn't seen these observations yet)
- Has one problem:

A.10 Hyperparameter

- Hyperparameter - parameter in the model that controls flexibility.
- e.g. Polynomial Regression $\rightarrow d$.
- Use Cross-Validation within the training set to tune the hyperparameter.
- Tuning Set, Training Set, Validation Set, and Testing Set

A.11 k-fold Cross Validation

- Usually $k = 5$ or $k = 10$. We use $k = 5$ in this class.
- Randomly assign data into $k + 1$ groups.
- For example, if $n = 155$ and $k = 5$,

```
n=155
[---------Tuning Set 125-------------]   [Test Set]
[fold 1][fold 2][fold 3][fold 4][fold 5]
[  25  ][  25  ][  25  ][  25  ][  25  ]   [   30   ]
```

A.12 k-fold Cross Validation

- Round 1
```
[------Training Set 100--------]   [validation set 25]
[fold 2][fold 3][fold 4][fold 5]   [fold 1]
[  25  ][  25  ][  25  ][  25  ]   [  25  ]
```

- Round 2
```
[------Training Set 100--------]   [validation set 25]
[fold 1][fold 3][fold 4][fold 5]   [fold 2]
[  25  ][  25  ][  25  ][  25  ]   [  25  ]
```

- Round 3
```
[------Training Set 100--------]   [validation set 25]
```

```
[fold 1][fold 2][fold 4][fold 5]    [fold 3]
[  25  ][  25  ][  25  ][  25  ]    [  25  ]
```

- Round 4

```
[------Training Set 100--------]    [validation set 25]
[fold 1][fold 2][fold 3][fold 5]    [fold 4]
[  25  ][  25  ][  25  ][  25  ]    [  25  ]
```

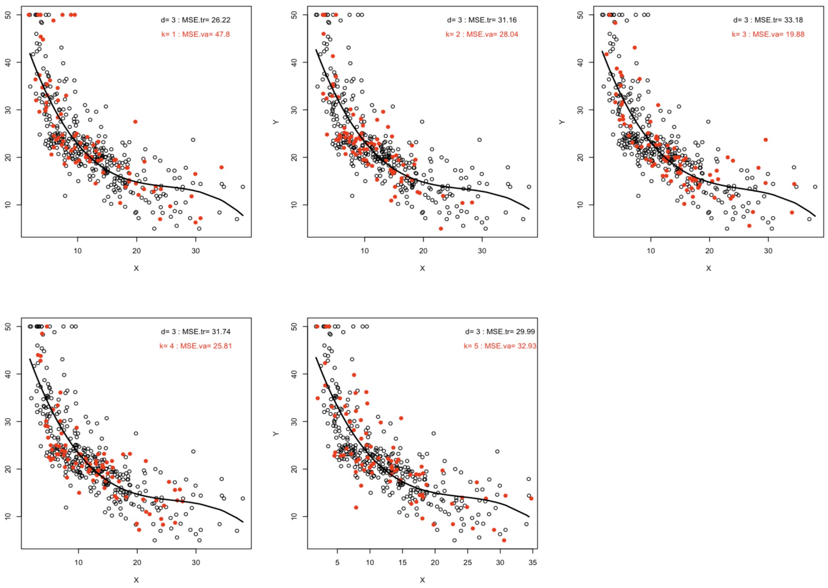
- Round 5

```
[------Training Set 100--------]    [validation set 25]
[fold 1][fold 2][fold 3][fold 4]    [fold 5]
[  25  ][  25  ][  25  ][  25  ]    [  25  ]
```

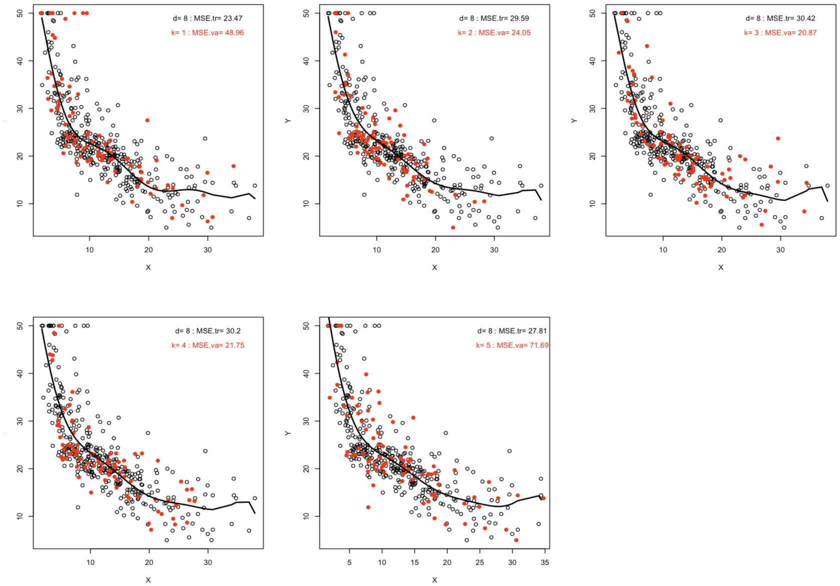
A.13 k-fold Cross Validation

- Take [Tuning Set], and use 5-fold CV and fit 5 times using 5 different [Training Set] and [Validation Set]
- Use average validation MSE to decide on the best value of the hyperpamameter.
- Now use the chosen hyperparameter, and fit entire [Tuning Set]. Then test it on [Test Set].
- Test Set should be used only once per method.
- Test MSE is the measure of performance for the method.

A.14 5-fold CV

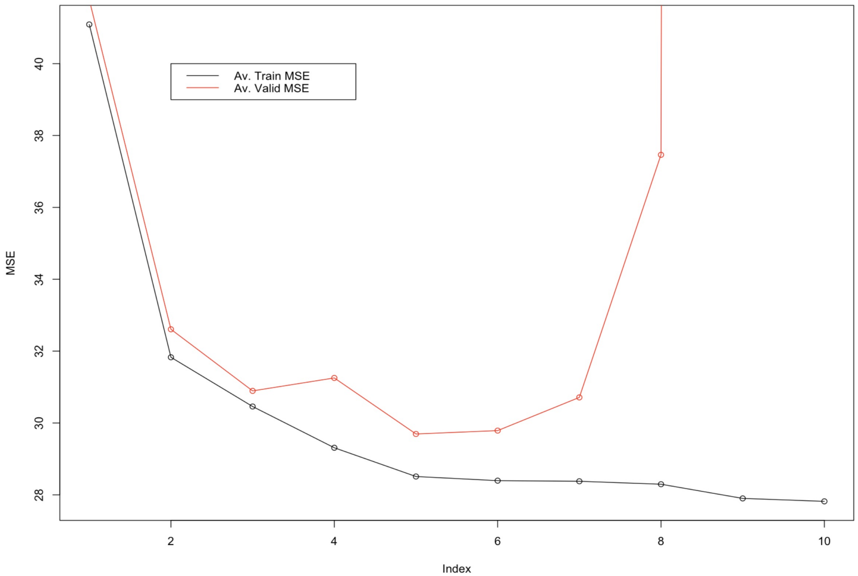


d= 3 : MSE.tr= 26.22
k= 1 : MSE.va= 47.8
d= 3 : MSE.tr= 31.16
k= 2 : MSE.va= 28.04
d= 3 : MSE.tr= 33.18
k= 3 : MSE.va= 19.88
d= 3 : MSE.tr= 31.74
k= 4 : MSE.va= 25.81
d= 3 : MSE.tr= 29.99
k= 5 : MSE.va= 32.93
X
Y

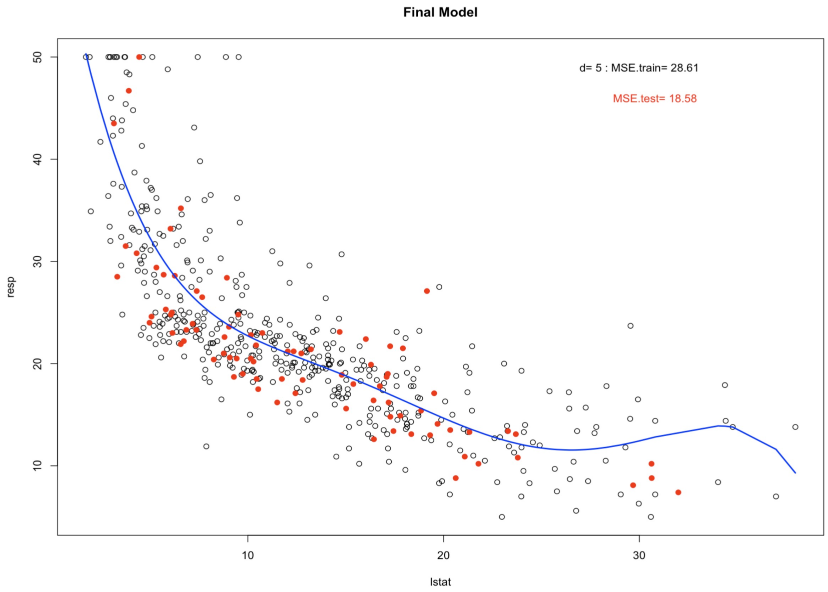


d= 8 : MSE.tr= 23.47
k= 1 : MSE.va= 48.96
d= 8 : MSE.tr= 29.59
k= 2 : MSE.va= 24.05
d= 8 : MSE.tr= 30.42
k= 3 : MSE.va= 20.87
d= 8 : MSE.tr= 30.2
k= 4 : MSE.va= 21.75
d= 8 : MSE.tr= 27.81
k= 5 : MSE.va= 71.69
X
Y

A.15 Training MSE vs Validation MSE



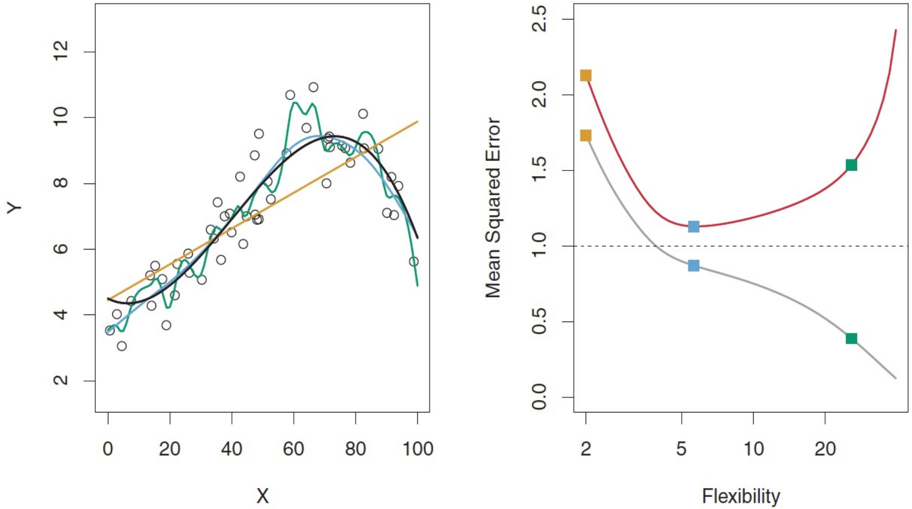
A.16 Final Test Fit



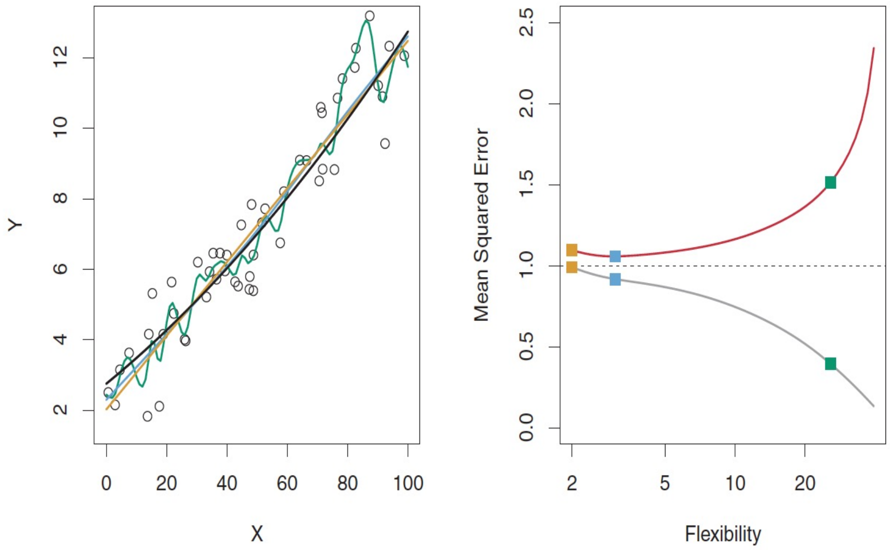
A.17 Bias-Variance Trade-Off

Prediction MSE can be decomposed as

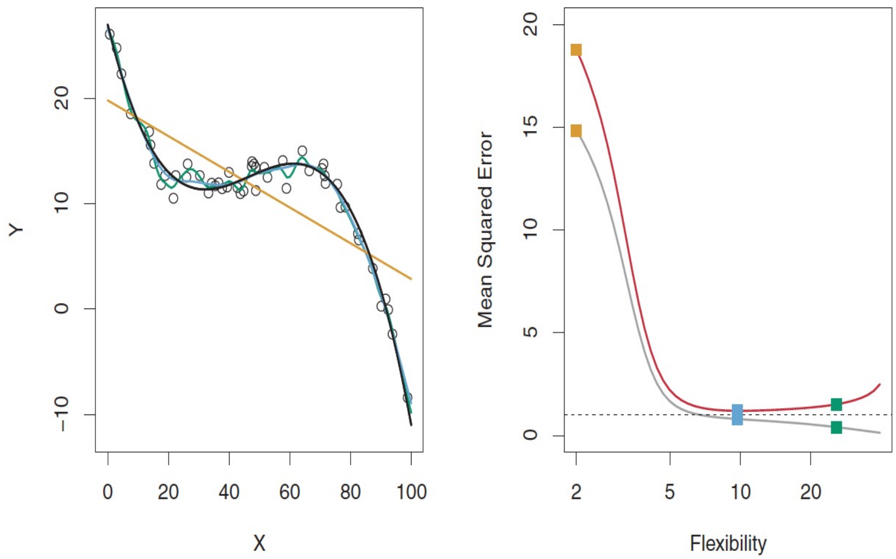
$$\begin{aligned}
E\big(Y - \hat{f}(X)\big)^2 &= E\Big(f(X) + \epsilon - \hat{f}(X)\Big)^2 \\
&= E\Big(f(X) - E(\hat{f}(X)) + E(\hat{f}(X)) - \hat{f}(X) + \epsilon\Big)^2 \\
&= E\Big(f(X) - E(\hat{f}(X))\Big)^2 + E\Big(E(\hat{f}(X)) - \hat{f}(X)\Big)^2 + E\big(\epsilon^2\big) \\
&= \Big(f(X) - E(\hat{f}(X))\Big)^2 + E\Big(E(\hat{f}(X)) - \hat{f}(X)\Big)^2 + E\big(\epsilon^2\big) \\
&= \Big[Bias(\hat{f}(X))\Big]^2 + Var(\hat{f}(X)) + Var(\epsilon)
\end{aligned}$$

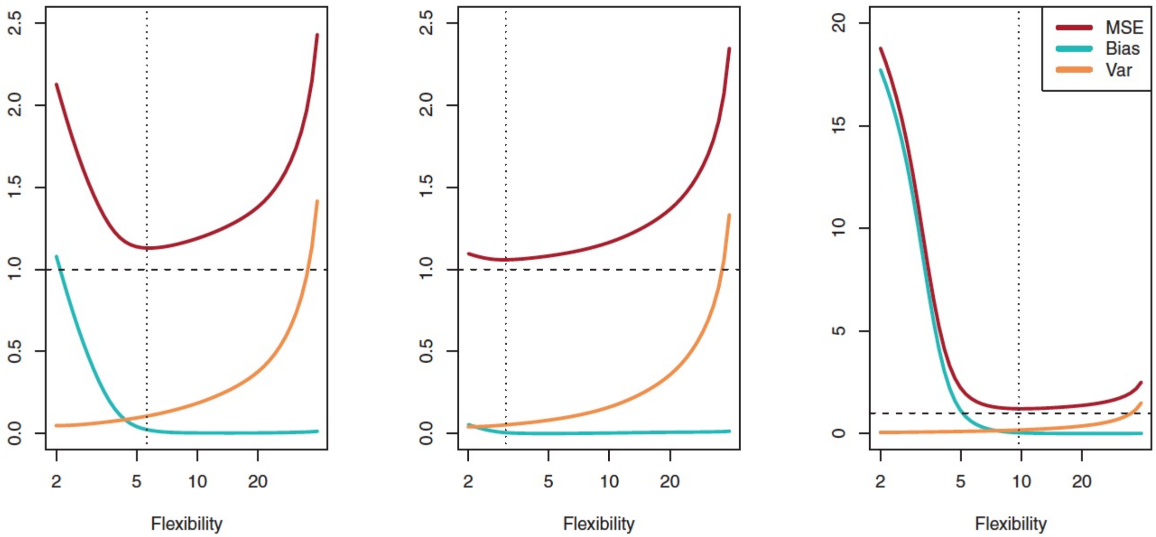


Y
X
Mean Squared Error
Flexibility



Y
X
0
20
40
60
80
100
2
4
6
8
10
12
Mean Squared Error
Flexibility
2
5
10
20
0.0
0.5
1.0
1.5
2.0
2.5


Y
20
10
0
−10
0
20
40
60
80
100
X
Mean Squared Error
20
15
10
5
0
2
5
10
20
Flexibility


MSE
Bias
Var
Flexibility
Flexibility
Flexibility

A.18 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

A.19 Assessing Model Prediction Accuracy

- If the model is fitting well, your

  $$\begin{aligned} &\text{[Av. Training MSE]} \\ &\text{[AV. Validation MSE]} \\ &\text{[Test MSE]} \end{aligned}$$

  should be all comparable.

- Your [Test MSE] is the best estimate for the true Prediction MSE.

A.20 In the Classification Setting

- Instead of MSE, work with Errror Rate:

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

A.21 Trade-off in the new approach

- Classical Statistics (Probabilistic Model)

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

  - Assume parametric model for $f(\cdot)$ and ϵ.
  - Sampling Probability of $(y_1, \cdots, y_n)$, which are realizations of r.v. Y.
  - Esimate parameters for $f(\cdot)$ and ϵ.
  - Because the model distinguish the mechanism $f(\cdot)$ vs noise ϵ, looking at in-sample fit was enough (if the assumption is correct).
  - Predict future Y using the estimated model.

- Pros and Cons
  - Model is interpretable.
  - Future effect of the model is easier to calculate.
  - No need for out-sample validation (test set), if assumption is correct.
  - Popular models are mathematically optimized already, to save the computational task.
  - Theory on prediction interval. Based on the assumption, often distribution on the prediction error is available.

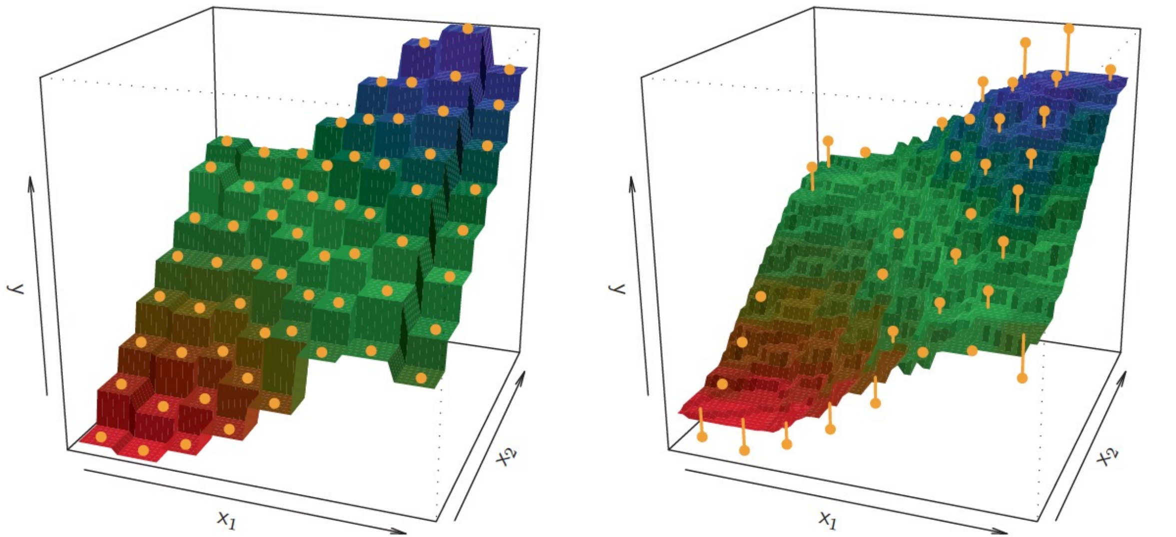
A.22 K-Nearest Neighbor

- One of elementary supervised learning model.
- Pick a point x_0, find K nearest observations.
- $f(x_0)$ is estimated by the average of all K neighbors:

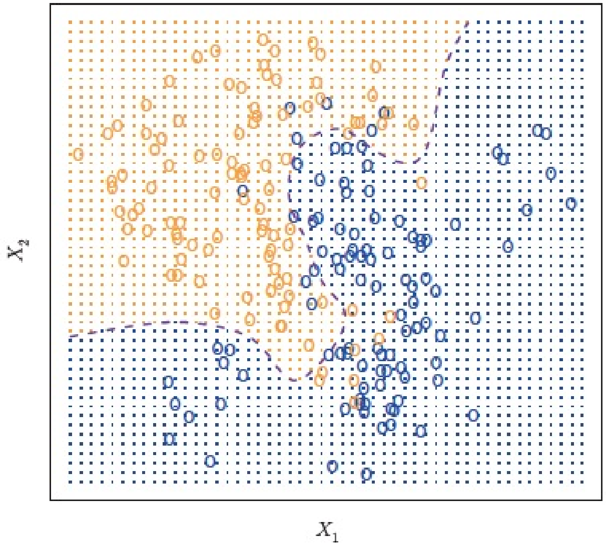
$$\hat{f}(x_0) = \frac{1}{K} \sum y_i.$$

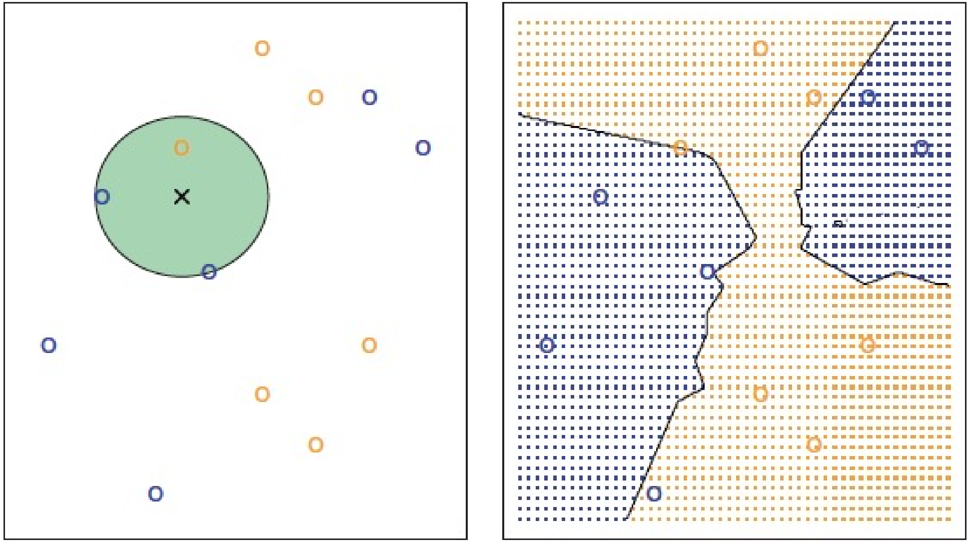
- K is the hyperparameter.
- Can be used for Regression or Classification

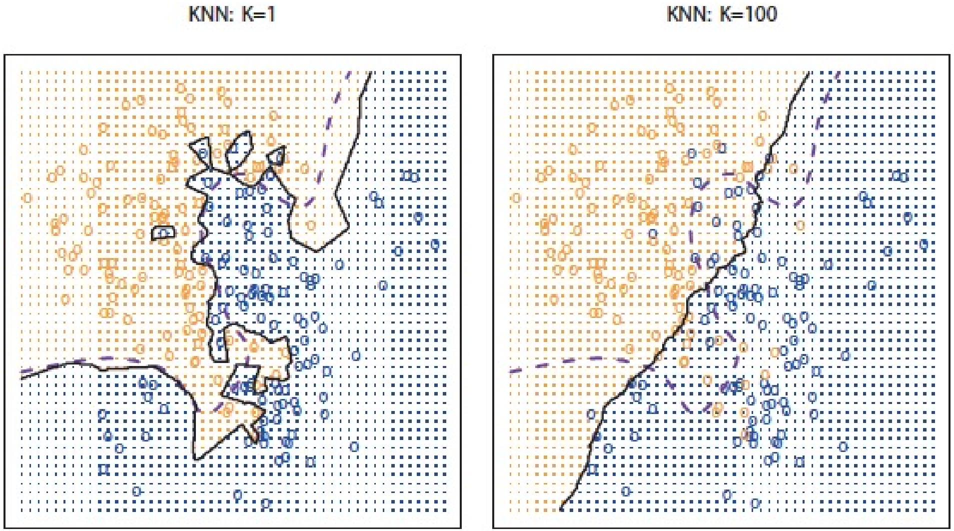
K=1 (left) and K=9 (right)

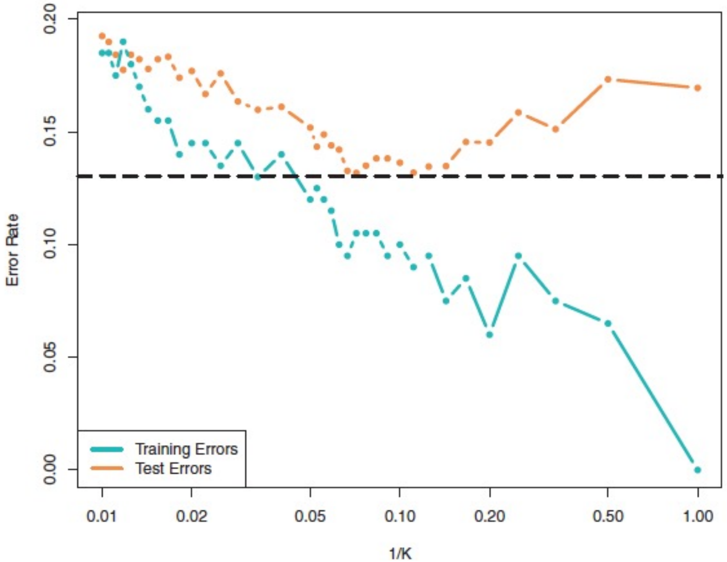


A.23 K-NN examples









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