(1)

(2)

As we can’t measure the error term ǫ. using Yˆ

to predict Y is simply impossible.

Therefore, we need to do some transformation. This is the so-called Population Regression Function.

E(Y ) = E(f (X) + ǫ) = f (X) (3)

Yˆ = fˆ(X) P−re→dict E(Y ) = f (x) (4)

Note that ǫ

Notation

i.∼i.d. N (0, σ2). So, E(ǫ) = 0.

* Y : the quantitative response of Y
* f (X): our model itself, where X = (X1, X2, X3, . . . , Xp) with p different predictors
* ǫ: Random error term, in linear regression ǫ

i.i.d.

∼ N (0, σ ) as an assumption

2

# Assessing Model Accuracy

* + 1. Measuring the Quality of Fit

In regression, the most common method to measure the fit of the data is the mean squared error(MSE), given by

n

1

M SE =

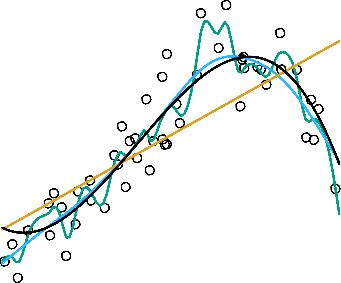
.(yi − fˆ(xi))2 (5)

n

i=1

* + - * Concept: Measuring the fit of training data
      * Goal: We are really not interested in whether fˆ(xi) ≈ yi; instead we want to know whether fˆ(x0) ≈ y0, where (x0, y0) is an unseen test observation not used to train the model.
    1. The Bias-Variance Trade-Off

linear regression model smoothing spline(p = 5) smoothing spline(p = 20)



testing MSE training MSE



Figure 1: The black curve represents the true model f . The other curves are fˆ which are used to predict E(f ).

As shown in the figure above, when the flexibility grows the MSE of the testing data increases dramatically and turns into a U-curve (curve of dimensionality). The orange line is the linear regression fit, which is relatively inflexible. In the left hand panel of Figure 1, the blue and green curves are smoothing splines with different levels of flexibility. As the model gets more flexible, it soon becomes over-fitted, making the MSE of test data increase.

The U-Shape observed in the test MSE curves turns out to be the result of two competing properties of statistical learning methods. The MSE for a given x0, can always be decomposed

into the sum of three fundamental quantities: the variance of fˆ(x0), the square Bias of fˆ(x0) and the variance for the error term ǫ. That is,

E(y0 − fˆ(x0))2 = E(y2) − 2E(y0)E(fˆ(x0)) + E[fˆ(x0)2]

0

= E(f (x0)2 + 2ǫf (x0) + ǫ2) − 2y0E(fˆ(x0)) + E(fˆ(x0)2)

= f (x0)2 + E(ǫ2) − 2y0E(fˆ(x0)) + [E(fˆ(x0)2) − E[fˆ(x0)]2] + E[fˆ(x0)]2

= f (x0)2 − 2y0E(fˆ(x0)) + E[fˆ(x0)]2 + V ar(fˆ(x0)) + V ar(ǫ)

= [E(fˆ(x0)) − f (x0)]2 + V ar(fˆ(x0)) + V ar(ǫ)

= [Bias(fˆ(x0))]2 + V ar(fˆ(x0)) + V ar(ǫ)

Unbiased Estimator and Bias

—————————————–

predict

θˆ

−→ θ

E(θˆ) = θ

Bias(θˆ) = E(θˆ) − θ

—————————————–

predict

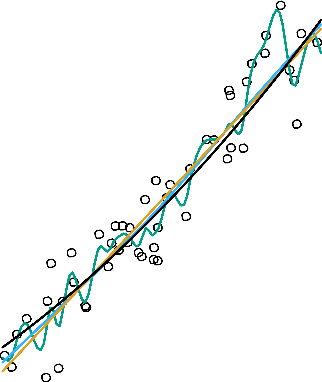
fˆ(x)

−→ f (x)

E(fˆ(x)) = f (x)

Bias(fˆ(x)) = E(fˆ(x)) − f (x)

linear regression model smoothing spline(p = 3) smoothing spline(p = 22)



Y

8

10 12

Mean Squared Error

1.5

2.0

2.5

testing MSE training MSE

4

6

0.5

1.0

0 20 40 60 80 100

2

0.0

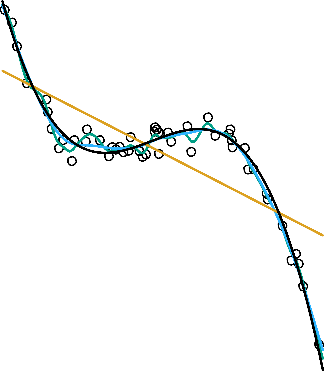
2 5 10 20

X Flexibility

Figure 2: Another f which is closer to linear. The black curve in the left panel is the true

f (X). In this setting, linear regression provides a good fit to the data.

linear regression model smoothing spline(p = 10) smoothing spline(p = 22)



10

20

Mean Squared Error

15

20

testing MSE training MSE

Y

0

5

10

0 20 40 60 80 100

−10

0

2 5 10 20

X Flexibility

Figure 3: Another f which is far from linear. In this setting, linear regression provides a poor fit to the data.

* + - * Variance : refers to the amount by which fˆ would change if we estimated it using a different training data set.
      * Bias: refers to the error that is introduced by approximating f

As a general rule, the more flexible a model gets, the more the variance increases and the more the bias decreases.

Figure 1 Figure 2 Figure 3

MSE

2.5

2.5

20

Bias Var

1.0

1.5

2.0

1.0

1.5

2.0

10

15

2 5 10 20

0.0

0.5

0.0

0.5

0

5

2 5 10 20

2 5 10 20

Flexibility

Flexibility

Flexibility

Figure 4: Square Bias(Blue Curve), Variance(Orange Curve) and MSE(Red Curve) from test data of Figure 1, Figure 2 and Figure 3.

Figure 4 represents the trade-off of variance and bias. Good statistical learning model requires low variance as well as low squared bias for the test data. However, in general, the bias decreases and the variance increases as the flexibility(dimension) of the model increases.

Most of the time in a real-life situation f is unknown. It is impossible to compute the test MES, bias, or the variance for a learning model. Nevertheless, one should always keep the bias-variance trade-off in mind.

* + 1. Classification Settings

Estimate f on the basis of training observations (x1, y1), . . . , (xn, yn), where y1, . . . , yn are

qualitative responses.

n

fˆ = 1 . I(y

fˆ is the training error rate, the proportion of mistakes :

ƒ= yˆ ) (6)

n i i

i=1

yˆi: the predicted class label for the ith observation

I(yi ƒ= yˆi) =



 1, yi ƒ= yˆi

 0, yi = yˆi

(7)

* + 1. The Bayes Classifier

More detail in Chapter 4.4

* + 1. Linear Models and Least Squares

Yˆ = XT βˆ

N

(8)

RSS(β) = .(yi − xT β)2 (9)

i

f.o.c

i=1

∂RSS(β)

∂β

= XT (y − Xβ) = 0 (10)

βˆ = (XT X)−1XT y (11)

R code:

x = matrix ( c ( 1 , 1 , 1 , 1 , 1 , 1 ,

0 . 3 5 , 0 . 4 7 5 , 0 . 5 6 , 0 . 5 4 , 0 . 6 1 , 0 . 5 9 ) , nrow = 6 , ncol = 2 )

y = matrix ( c ( 4 , 5 . 2 5 , 6 . 8 , 6 . 4 5 , 7 . 8 , 7 . 5 5 )

, nrow = 6 , ncol = 1 ) f i t 1 <− lm( y ˜ x )

solve ( ( t ( x ) %∗% x ) ) %∗% ( t ( x ) %∗% y )

* + 1. Nearest-Neighbour Methods

The model is defined as follows:

Yˆ (x) = 1

k

. yi

xi ∈Nk (x)

(12)

* + - * Concept : average k training points xi nearest to x
      * Nk : the neighbourhood of x defined by the k closest points xi in the training sample

T

* + - * xi: represent the training data in input space, i = 1, 2, 3, ..., N

Implementation:

* + - * Step 1: Choose data point x
      * Step 2: Find the k observations with xi closest(Euclidean Distance) to x in input space and average their response to get Yˆ (x)
      * Step 3: Move to the next data point and repeat above
    1. Nearest-Neighbour Methods: as a classifier

The model is defined as follows:

Pr(Y = j|X = x0) = K . I(yi = j) (13)

1

i∈N0

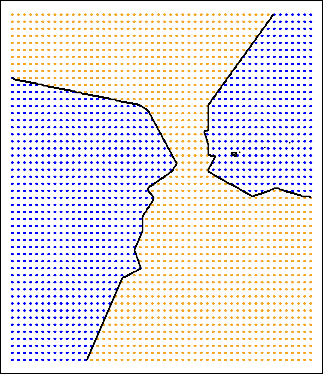
* + - * .

i∈N0

I(yi = j) : Summarize k points nearest to x0 according to class j

Example:

k-Nearest Neighbour (k = 3)

o o

o o o o

o o o o o o

o o

o o o o o o

o o

o o

x0



 1 .

I(yi = Blue), j = Blue





Pr(Y = j|X = x0) =



K

i∈N0

1

(14)

K .

i∈N0





I(yi = Orange), j = Orange

 0, otherwise



 1 .

I(yi = Blue) = 2 , j = Blue

3





Pr(Y = j|X = x0) =



3

i∈N0

1 1

(15)

3 .

i∈N0





I(yi = Orange) = 3 , j = Orange

 0, otherwise

Inference:

The probability for Blue circle (2/3) is higher than Orange circle (1/3). The test data

x0 should be classified as Blue circle.

# Exercises

* + 1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an

inflexible method. Justify your answer.

* + - 1. The sample size n is extremely large, and the number of predictors p is small.
      2. The number of predictors p is extremely large, and the number of observations n

is small.

* + - 1. The relationship between the predictors and response is highly non-linear.
      2. The variance of the error terms, i.e. σ2 = Var(ǫ), is extremely high.

Answer:

1. Better. If we have sufficient data, then it’s better to fit a flexible model, since we can picture our true model easily with large dataset. ~~Worse. When a sample data~~ ~~is extremely large and the predictors are few. It’s better to use a inflexible model,~~ ~~since lack of predictors will more fit with inflexible model.~~
2. Worse. If we fit a flexible model with the number of observations is small, it’ll course over-fitting. ~~Better. When predictors are extremely large and data are few.~~ ~~It’s better to use a flexible model, since large predictors will more fit with flexible~~ ~~model.~~
3. Better. A non-linear model consists of high dimension. This property makes the model more flexible.
4. Worse. When a model with a high σ2, then it’s better to apply an inflexible model to control the variance.
   * 1. Explain whether each scenario is a classification or regression problem, and indi- cate whether we are most interested in inference or prediction. Finally, provide n and p.
        1. We collect a set of data on the top 500 firms in the US. For each firm we record

profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

* + - 1. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
      2. We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

Answer:

1. • Regression problem, since we want to know which predicator effect the CEO salary the most.
   * Inference

• n = 500

* + p:profit, number of employees and industry
  + Response: the CEO salary

1. • Classification problem, since we make use of the previous data to predict whether the new product will success of not.
   * Prediction

• n = 20

* + p: price charged for the product, marketing budget, competition price, and ten other variables
* Response: success or failure

1. • Regression problem, since we want to predict % change in the US dollar
   * Prediction

• n = 52

* + p: the % change in the US dollar, the % change in the British market, and the % change in the German market.
  + Response: % change in the US dollar

1. The table below provides a training data set containing six observations, three predic- tors, and one qualitative response variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Obs. | X1 | X2 | X3 | Y |
| 1 | 0 | 3 | 0 | Red |
| 2 | 2 | 0 | 0 | Red |
| 3 | 0 | 1 | 3 | Red |
| 4 | 0 | 1 | 2 | Green |
| 5 | -1 | 0 | 1 | Green |
| 6 | 1 | 1 | 1 | Red |

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 =

X3 = 0 using K-nearest neighbours.

* 1. Compute the Euclidean distance between each observation and the test point,

X1 = X2 = X3 = 0

* 1. What is our prediction with K = 1? Why?
  2. What is our prediction with K = 3? Why?
  3. If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

Answer:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Obs. | X1 | X2 | X3 | Y | Euclidean |
| x0 | 0 | 0 | 0 |  |  |
| 1 | 0 | 3 | 0 | Red | 3 |
| 2 | 2 | 0 | 0 | Red | 2 |
| 3 | 0 | 1 | 3 | Red | 3.16 |
| 4 | 0 | 1 | 2 | Green | 2.23 |
| 5 | -1 | 0 | 1 | Green | 1.41 |
| 6 | 1 | 1 | 1 | Red | 1.73 |

(a)

1. If K=1, then the category of obs.5 will be the most similar one, since the closest distance. The prediction of fˆ(x0) will be Green
2. If K=3, we’ll choose the top 3 data points which are closest to x0. In this case, obs.5, obs.6 and obs.2 will be selected. The probability can be illustrated as below:









Pr(Y = j|X = x0) =







3 . i∈(5,6,2) 3 .

1

1

i∈(5,6,2)

I(yi = Red) = 2 , j = Red

3

I(yi = Green) = 1 , j = Green

3

(16)

 0, otherwise

Since, x0 is more likely to be Red according to higher probability ( 2 ). We’ll predict

3

x0 as Red.

1. If the problem is highly non-linear, K should be small. Due to KNN’s property, when k is small it means the model is more flexible whereas a large K would try to fit a more linear boundary because it takes more points into consideration.