2 Statistical Learning

2.1 Model

Often, our model can be illustrated as below

$$Y = f(X) + \epsilon \tag{1}$$

Most of the time we can't get the true model Y because the parameters for the predicator in $f(X)$ are unknown. In this case, we need a prediction model to predict the response Y.

$$\hat{Y} = \hat{f}(X) \tag{2}$$

As we can't measure the error term ϵ, using $\hat{Y}$ to predict Y is simply impossible. Therefore, we need to do some transformation. This is so-called Population Regression Function.

$$E(Y) = E(f(X) + \epsilon) = f(X) \tag{3}$$

$$\hat{Y} = \hat{f}(X) \overset{Predict}{\longrightarrow} E(Y) = f(x) \tag{4}$$

Note that $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. So, $E(\epsilon) = 0$.

Notation

- Y: the quantitative response of Y
- $f(X)$: our model itself, where $X = (X_1, X_2, X_3, \ldots, X_p)$ with p different predictors
- ϵ: Random error term, in linear regression $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ as an assumption

2.2 Assessing Model Accuracy

2.2.1 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

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{5}$$

- Concept: Measuring the fit of training data
- Goal: We are really not interested in whether $\hat{f}(x_i) \approx y_i$; instead we want to know whether $\hat{f}(x_0) \approx y_0$, where (x_0, y_0) is an unseen test observation not used to train the model.

2.2.2 The Bias-Variance Trade-Off

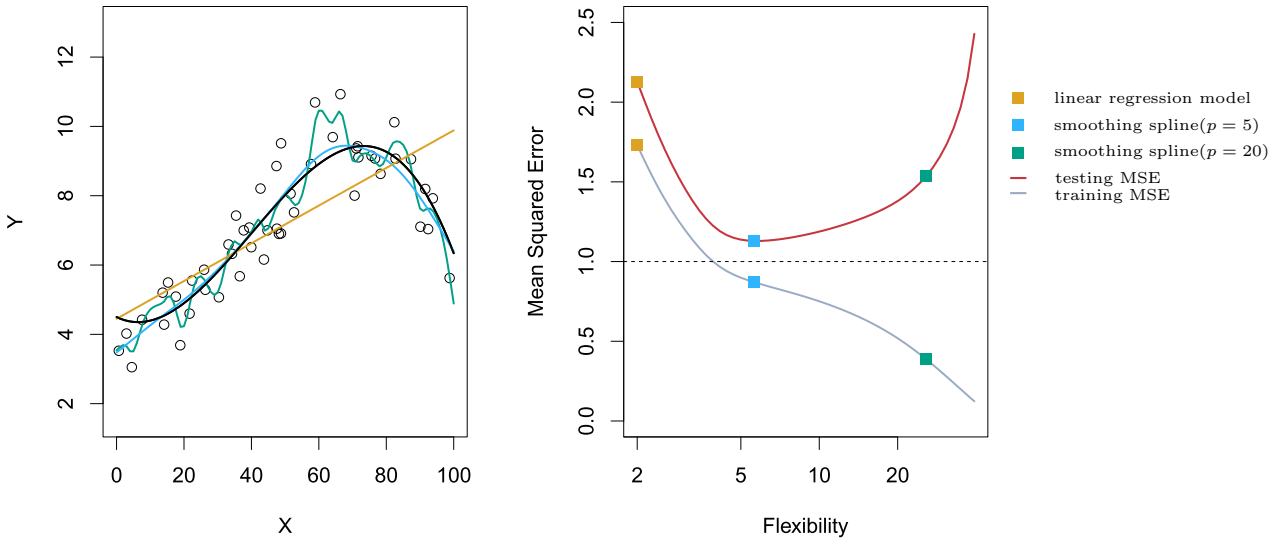


Figure 1: The black curve represents the true model f. The other curves are $\hat{f}$ which are used to predict $E(f)$.

As shown in the figure above, when the flexibility grows the MSE of the testing data increases dramatically and turns into a U-curve. 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 x_0, can always be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the square Bias of $\hat{f}(x_0)$

and the variance for the error term ϵ. That is,

$$\begin{aligned}
E(y_0 - \hat{f}(x_0))^2 &= E(y_0^2) - 2E(y_0)E(\hat{f}(x_0)) + E[\hat{f}(x_0)^2] \\
&= E(f(x_0)^2 + 2\epsilon f(x_0) + \epsilon^2) - 2y_0 E(\hat{f}(x_0)) + E(\hat{f}(x_0)^2) \\
&= f(x_0)^2 + E(\epsilon^2) - 2y_0 E(\hat{f}(x_0)) + \underline{[E(\hat{f}(x_0)^2) - E[\hat{f}(x_0)]^2]} + E[\hat{f}(x_0)]^2 \\
&= f(x_0)^2 - 2y_0 E(\hat{f}(x_0)) + E[\hat{f}(x_0)]^2 + Var(\hat{f}(x_0)) + Var(\epsilon) \\
&= [E(\hat{f}(x_0)) - f(x_0)]^2 + Var(\hat{f}(x_0)) + Var(\epsilon) \\
&= [Bias(\hat{f}(x_0))]^2 + Var(\hat{f}(x_0)) + Var(\epsilon)
\end{aligned}$$

Unbiased Estimator and Bias

$\hat{\theta} \overset{estimate}{\longrightarrow} \theta$

$E(\hat{\theta}) = \theta$

$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$

$\hat{f}(x) \overset{estimate}{\longrightarrow} f(x)$

$E(\hat{f}(x)) = f(x)$

$\text{Bias}(\hat{f}(x)) = E(\hat{f}(x)) - f(x)$

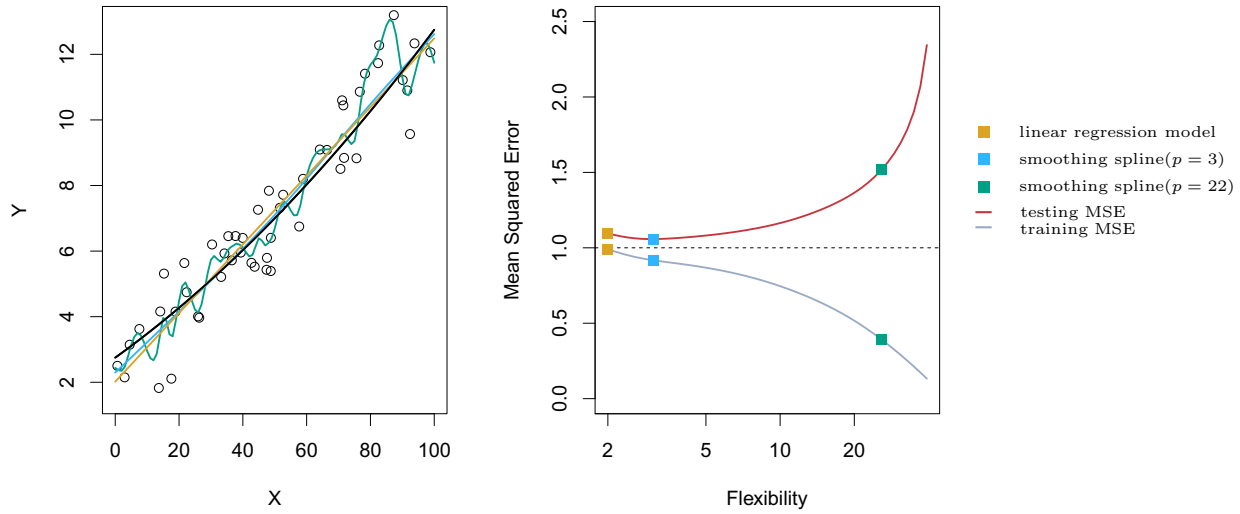


Figure 2: Anthother f which is closer to linear. The black curve in the left panel is the true $f(X)$. In this setting, linear regression provides a good fit to the data.

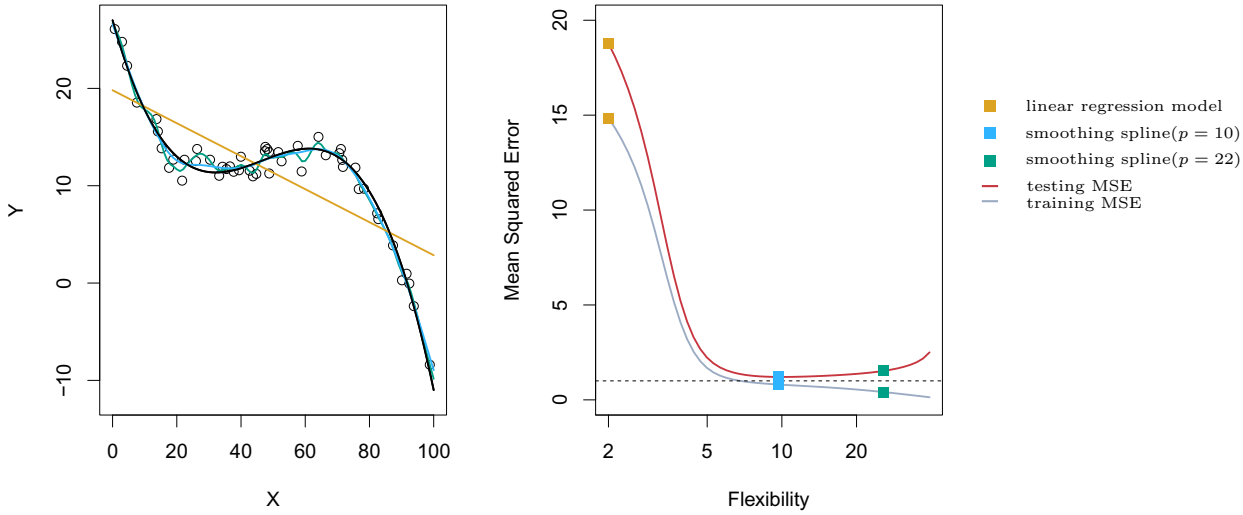


Figure 3: Anthother f which is far from linear. In this setting, linear regression provides a poor fit to the data.

- Variance : refers to the amount by which $\hat{f}$ 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 models gets, the variance increases and the more the bias decrease.

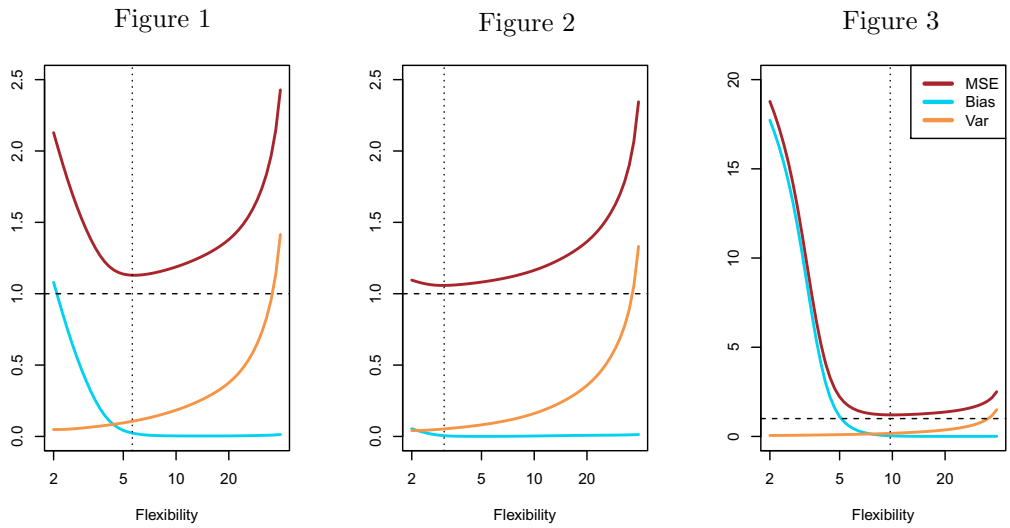


Figure 4: Square Bias(Blue Curve), Variance(Orange Curve) and MSE(Red Curve) from test data of Figure 1, Figure 2 and Figure 3.

The 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.

2.2.3 Classification Settings

Estimate f on the basis of training observations $(x_1, y_1), \ldots, (x_n, y_n)$, where $y_1, \ldots, y_n$ are qualitative response. $\hat{f}$ is the training error rate, the proportion of mistakes :

$$\hat{f} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{6}$$

$\hat{y}_i$: the predicted class label for the ith observation

$$I(y_i \neq \hat{y}_i) = \begin{cases} 1, & y_i \neq \hat{y}_i \\ 0, & y_i = \hat{y}_i \end{cases} \tag{7}$$

2.2.4 The Bayes Classifier

More detail in Chapter 4.4

2.2.5 Linear Models and Least Squares

$$\hat{Y} = X^T\hat{\beta} \tag{8}$$

$$\text{RSS}(\beta) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2 \tag{9}$$

$$\text{f.o.c} \quad \frac{\partial \text{RSS}(\beta)}{\partial \beta} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0 \tag{10}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{11}$$

R code:

```
x = matrix(c  (1,1,1,1,1,1,0.35,0.475,0.56,0.54,0.61,0.59),  nrow = 6, ncol = 2)
y = matrix(c  (4,5.25,6.8,6.45,7.8,7.55),  nrow = 6, ncol = 1)
fit1  = lm(y ~ x)
solve((t(x) %*% x)) %*% (t(x) %*% y)
```

2.2.6 Nearest-Neighbour Methods: to predict a quantitative response

The models is defines as follows:

$$\hat{Y}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i \tag{12}$$

- Concept : average k training points x_i nearest to x

- N_k: the neighbourhood of x defined by the k closest points x_i in the training sample $\mathcal{T}$
- x_i: 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 x_i closest(Euclidean Distance) to x in input space and average their response to get $\hat{Y}(x)$
- Step 3: Move to the next data point and repeat above

2.2.7 Nearest-Neighbour Methods: to predict a qualitative response

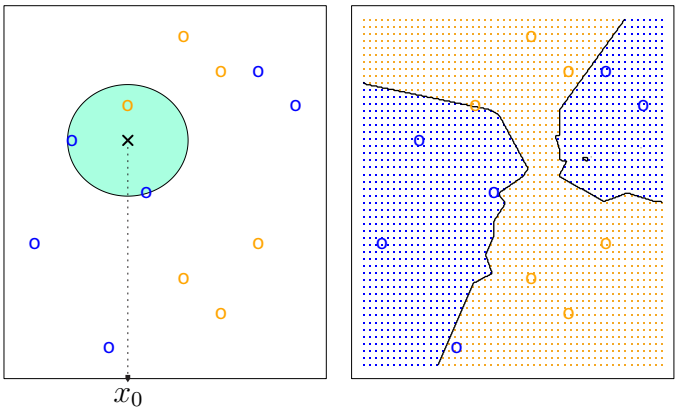
The models is defines as follows:

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) \tag{13}$$

- $\sum_{i \in N_0} I(y_i = j)$: Summarize k points nearest to x_0 according to class j

Example:

k-Nearest Neighbour ($k = 3$)



$$\Pr(Y=j|X=x_0)=\begin{cases}\frac{1}{K}\sum_{i\in N_0} I(y_i=\text{Blue}), & j=\text{Blue}\\ \frac{1}{K}\sum_{i\in N_0} I(y_i=\text{Orange}), & j=\text{Orange}\\ 0, & \text{otherwise}\end{cases} \tag{14}$$

$$\Pr(Y=j|X=x_0)=\begin{cases}\frac{1}{3}\sum_{i\in N_0} I(y_i=\text{Blue})=\frac{2}{3}, & j=\text{Blue}\\ \frac{1}{3}\sum_{i\in N_0} I(y_i=\text{Orange})=\frac{1}{3}, & j=\text{Orange}\\ 0, & \text{otherwise}\end{cases} \tag{15}$$

Inference:

The probability for Blue circle (2/3) is higher than Orange circle (1/3). The test data x_0 should be classified as Blue circle.

2.3 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 <u>better or worse</u> than an inflexible method. Justify your answer.

   (a) The sample size n is extremely large, and the number of predictors p is small.

   (b) The number of predictors p is extremely large, and the number of observations n is small.

   (c) The relationship between the predictors and response is highly non-linear.

   (d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

Answer:

(a) 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.

(b) Worse. If we fit a flexible model with the number of observations is small, it'll course over-fitting.

(c) Better. A non-linear model consists of high dimension. This property makes the model more flexible.

(d) Worse. When a model with a high σ^2, then it's better to apply a inflexible model to control the variance.

2. Explain whether each scenario is a **classification or regression** problem, and indicate whether we are most interested in **inference or prediction**. Finally, provide n **and** p.

   (a) 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.

   (b) 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.

   (c) 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:

(a)
- 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

(b)
- 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

(c)
- 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

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	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 $X_1 = X_2 = X_3 = 0$ using K-nearest neighbours.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$

(b) What is our prediction with K = 1? Why?

(c) What is our prediction with K = 3? Why?

(d) 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:

(a)

Obs.	X_1	X_2	X_3	Y	Euclidean
x_0	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

(b) If K=1, then the category of obs.5 will be the most similar one, since the closest distance. The prediction of $\hat{f}(x_0)$ will be Green

(c) If K=3, we'll choose the top 3 data points which are closest to x_0. In this case, obs.5, obs.6 and obs.2 will be selected. The probability can be illustrated as below:

$$\Pr(Y = j|X = x_0) = \begin{cases} \frac{1}{3} \sum_{i \in (5,6,2)} I(y_i = \text{Red}) = \frac{2}{3}, & j = \text{Red} \\ \frac{1}{3} \sum_{i \in (5,6,2)} I(y_i = \text{Green}) = \frac{1}{3}, & j = \text{Green} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Since, x_0 is more likely to be Red according to higher probability ($\frac{2}{3}$). We'll predict x_0 as Red.

(d) 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.

Reference

張翔 (2012), 《提綱契領學統計》, 四版, 鼎茂圖書

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.