MATH4432 Notes

SmokingPuddle58

September 26, 2024

This is the lecture note typed by SmokingPuddle in September, 2024. It mainly contains what professor mentions starting from year 3. For the contents of the first two weeks, I will try my best to include as much as possible.

The main reference source comes from the professor himself, lecture notes, tutorial notes, and also from the Internet if necessary.

Please inform me if there is any errors, better within the semester or I will have a very high chance of forgetting the contents.

Theorems, Corollary, Lemma, Proposition

Definitions

Examples

Warnings / Remarks

Proofs, Answers

Some special symbols, notations and functions that will appear in this note:

$\mathbb{C}$	Set of complex numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{Z}$	Set of integers
$\mathbb{Q}$	Set of rational numbers

Contents

1 Overview

1.1 Introduction

Before we start, we shall clarify some of the notations that will be used.

Consider the following expression:

$$P(X = x)$$

If we say $r.v.$ (Random variable) X, we actually means the name of the variable, while for x, we means the realization for such $r.v.$

Suppose we are now observing some quantitative response Y and also input variable X, consisting of p features, which can be expressed as:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_p \end{bmatrix}$$

where $X_1, ..., X_p$ are random variables. Then the relation between Y and X can be expressed as:

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

where ε is the error term independent from X, with mean 0, and f is a deterministic function. We call such model the population level model, or ground truth model. (i.e. The number of samples is infinitely many)

Remark 1.1
Note that Y, ε are all random variables, while X is a collection of random variables.

If we want to consider a sample level (the realization of the random variables), then the equation becomes:

$$y_i = \quad f(x_i) + \varepsilon_i \quad i = 1, ..., n$$

where x_i can be a vector like the following:

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ x_{ip} \end{bmatrix}$$

and n is the sample size.

Remark 1.2
In machine learning, vectors usually means **column vectors, but not row vectors**.

For example, consider the equation $f(x) = a_1x_1 + a_2x_2 + a_3x_3$. If we know that $a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, then $f(x) = a^\intercal x$, where $a^\intercal$ is the transpose of the vector a.

Now let's go back to the ground truth model, which is $Y = f(X) + \varepsilon$. Suppose we want to construct f from the data, then we will have:

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

for any observed x.

Suppose we are interested in the difference between the data and the observed prediction, then we will be interested in the value of $\mathbb{E}(Y - \hat{Y})^2$, the expected square error.

Remark 1.3
Both Y and $\hat{Y}$ are random variable, since $\hat{Y}$ is the prediction that is learnt from the data, and data comes from the random sample chosen from ground truth model. Thus we are not interested in the value of $(Y - \hat{Y})^2$, since it is not fixed.

Theorem 1.1

$$\mathbb{E}(Y - \hat{Y})^2 = \underbrace{\mathbb{E}(f(X) - \hat{f}(X))^2}_{\text{Reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible}}$$

Proof.

$$\begin{aligned}
\mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}(f(X) + \varepsilon - \hat{f}(X))^2 \\
&= \mathbb{E}(f(X) - \hat{f}(X) + \varepsilon)^2 \\
&= \mathbb{E}((f(X) - \hat{f}(X))^2 + 2\varepsilon(f(X) - \hat{f}(X)) + \varepsilon^2) \\
&= \mathbb{E}((f(X) - \hat{f}(X))^2) + \mathbb{E}(2\varepsilon(f(X) - \hat{f}(X))) + \mathbb{E}(\varepsilon^2) \\
&= \mathbb{E}((f(X) - \hat{f}(X))^2) + \underbrace{\mathbb{E}(2\varepsilon)\mathbb{E}((f(X) - \hat{f}(X)))}_{\text{Assume } \varepsilon \text{ independent from } f, \hat{f}} + \mathbb{E}(\varepsilon^2) \\
&= \mathbb{E}(f(X) - \hat{f}(X))^2 + \underbrace{0}_{\mathbb{E}(\varepsilon)=0} + \mathbb{E}(\varepsilon^2)
\end{aligned}$$

Since for any random variable X, and its expected value, $E(X) = \mu$, we have: (Covered in MATH2411)

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}(X - \mu)^2 \\
&= \mathbb{E}(X^2) + \mathbb{E}(\mu^2) - 2\mathbb{E}(X\mu) \\
&= \mathbb{E}(X^2) + \mu^2 - 2\mu\mathbb{E}(X) \\
&= \mathbb{E}(X^2) + \mu^2 - 2\mu^2 \\
&= \mathbb{E}(X^2) - \mu^2
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}(f(X) - \hat{f}(X))^2 + \mathbb{E}(\varepsilon^2) \\
&= \mathbb{E}(f(X) - \hat{f}(X))^2 + \text{Var}(\varepsilon)
\end{aligned}$$

To conclude, you can only reduce the error for the reducible part, by making your model approximate the ground truth model as well as possible, while we can really do not much on the irreducible part.

1.2 Estimation of f

There are two methods for estimating f, namely parametric, and non-parametric methods.

For parametric method, we assume that f can be described by a set of parameters, such that once all of the parameters are known, then the model is known.

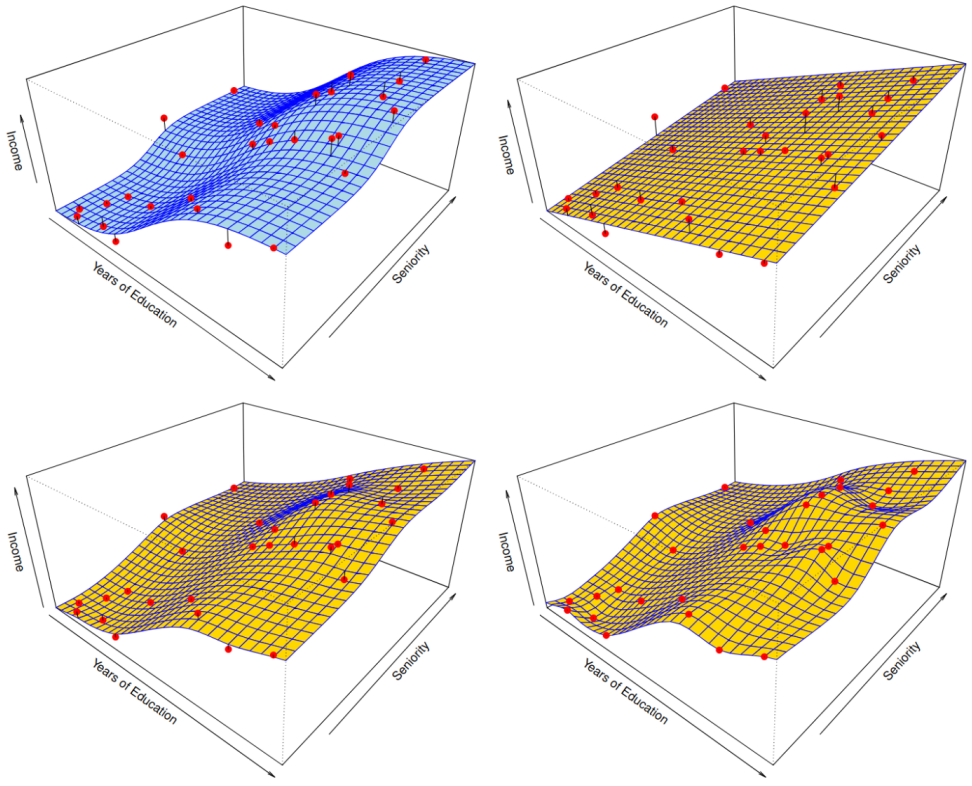
Example 1.1
One of the most simple assumption is that f is linear in X, which can be described as:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

For non-parametric method, we do not pre-specify the form of the model (Can be linear, non-linear. tree and neural network). To control the flexibility, we can always tune the parameters.

The advantages and disadvantages are listed in the following table.

	Parametric	Non-parametric
Advantage	Easy to solve and understand	More flexible and sometimes more powerful
Disadvantage	The model may be too simple to fit into the data	The model may be too flexible, there may be overfitting



Ground truth model	Underfitting
Good estimation	Overfitting (Fitting into the noise)

The above image shows the example of a ground truth model, a good estimation, overfitting and underfitting. It is also included in the lecture note.

1.3 Assessing Model Accuracy

In statistical machine learning context, to find a good estimate f, we shall introduce the concept of regularization (Covered in detail later), in which we want to minimize the following value as much as possible:

$$\text{Loss}(Y, f) + \lambda R(f)$$

where Y is the response, λ is a tuning parameter (weight) for the regularizer $R(f)$ of the model f.

The introduction of the regularizer is trying to control the complexity / flexibility of the function f to prevent perfect fit / overfitting issue.

Example 1.2 (Spline interpolation)
Consider the loss function and the regularizer as:

$$\underbrace{\sum_{i=1}^{n_{\text{train}}} (y_i - g(x_i))^2}_{\text{Loss function}} + \lambda \underbrace{\int g''(t)^2 \, \mathrm{d}t}_{\text{Regularizer}}$$

The regularizer tries to control the second derivative of g to be as small as possible to ensure smoothness.

Assume that λ is a extremely large value, the only solution is to let the integral to be 0, i.e. $g(t)$ should be a linear function.

If we put $\lambda = 0$, then the solution will be

$$\hat{g}(x_i) = y_i$$

i.e. The model will have a perfect fit to the data.

If we tune the parameter correctly, then we can get a good model that is neither overfit nor underfit.

So how do we know if our tuning parameter is good or not?

Consider there is a set of data $\{(x_i, y_i)\}_{i=1\ldots n}$, and we want to minimize

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda R(f)$$

as much as possible.

To achieve that, we need to spilt the data into two parts: Training data and Testing data, and the two set of datasets should not be overlapping with each other.

The idea is, we only use the training data for solving optimization. After such process, we will obtain $\hat{g}$, which is estimated from the training data. The testing data is then used to evaluate whether the model is accurate or not. i.e. We would like to evaluate the value of

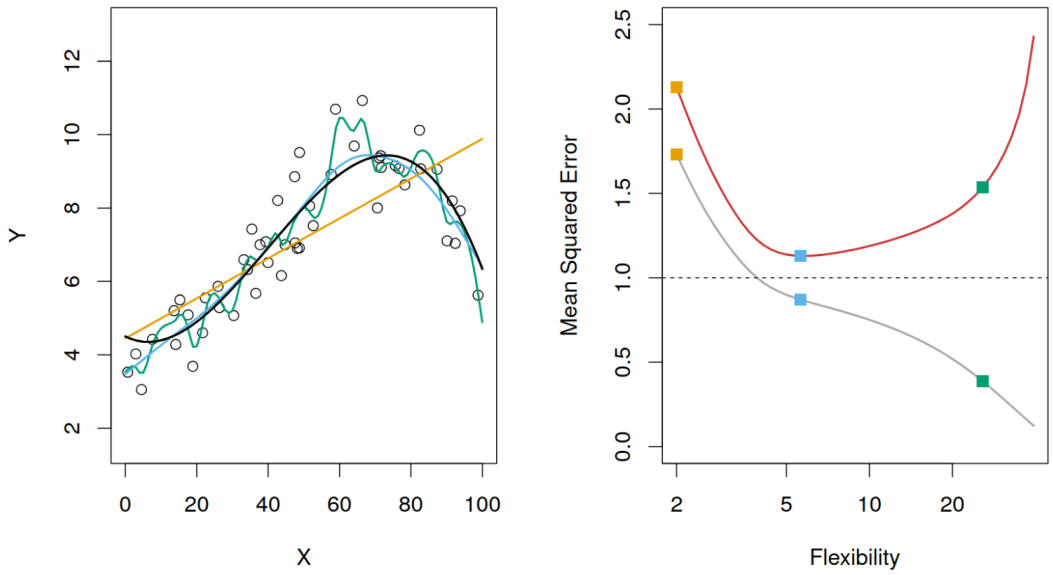
$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{g}(x_i))^2$$

where n is the number of testing data. Such value is defined as MSE (Mean Square Error).

Definition 1.1 (Mean Square Error)

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

where $g(x_i)$ is the prediction $\hat{g}$ gives for the $i-$th prediction.



Black: Ground truth model	Red: Testing error
Yellow: Simple linear model	Gray: Training error
Green: Perfect model (Fit all data into the model)	

In the red curve, there are three point corresponding to the three different models on the left, which shows the MSE of overfitting , underfitting and a good model.

Definition 1.2 (Training and Testing Error)
Suppose $\hat{g}_\lambda$ is estimated from the data, and depends on λ, then the accuracy of $\hat{g}_\lambda$ based on the testing data is given by:

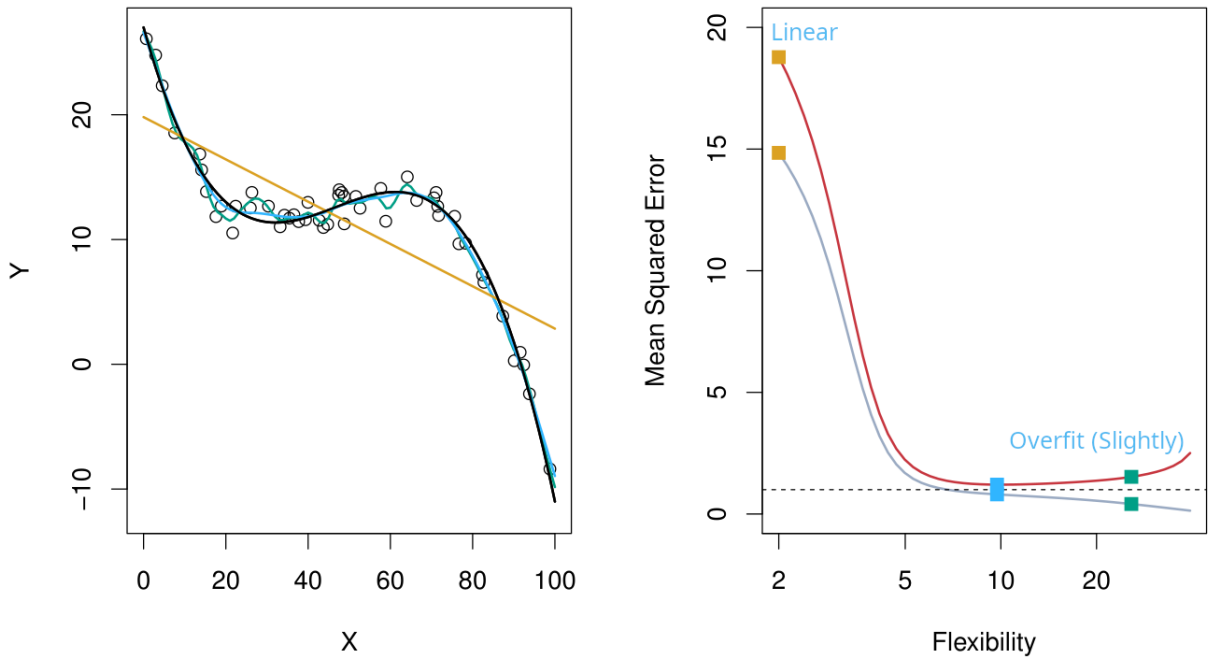
$$\frac{1}{n_{\text{Test}}} \sum_{i \in \text{Test}} (y_i - \hat{g}_\lambda(x_i))^2$$

This is called **testing mean square error (TMSE)**.

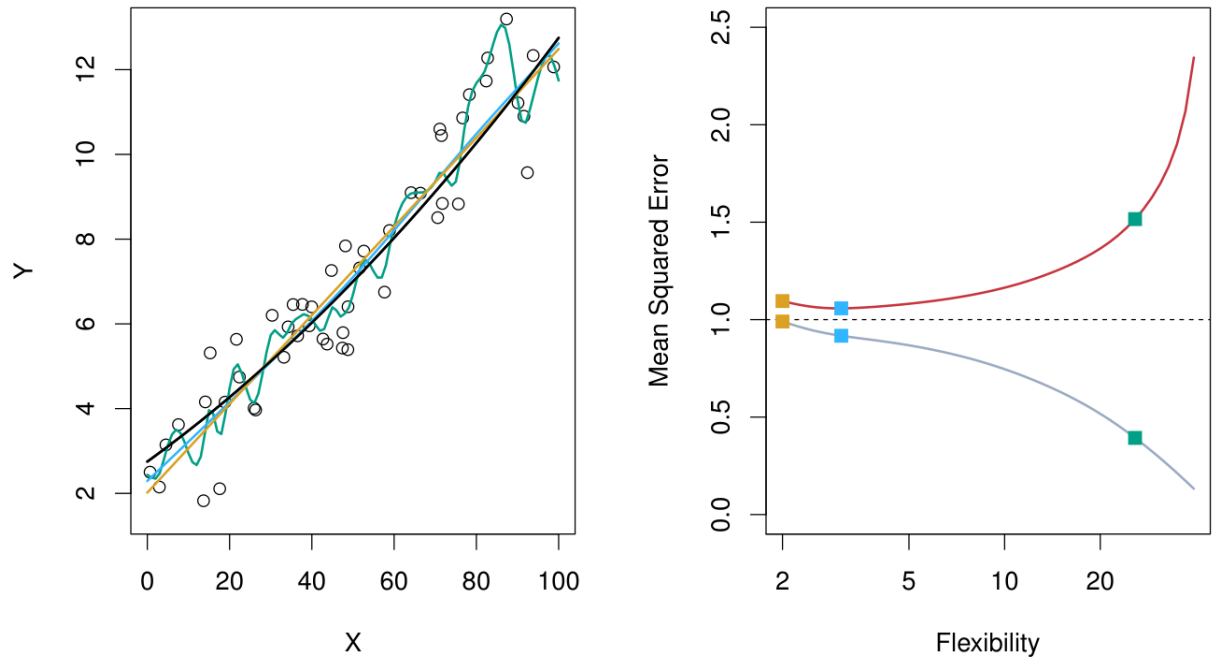
The **training error** is given by:

$$\frac{1}{n_{\text{Train}}} \sum_{i \in \text{Train}} (y_i - \hat{g}_\lambda(x_i))^2$$

Note 1
Training error is usually less than testing error, since optimization is solved with testing error, especially when you put $\lambda = 0$.



Note that linear regression above provides a very poor fit to the data, while perfect fit model does not increase error too much, since the noise is very small.



For this dataset, the ground truth model is close to linear, and due to the large noise, the high flexibility gives more error.

1.4 Bias-Variance Trade-Off

As shown in the previous example, we can see that the testing error are in U-shape. To understand the reason, we will need to introduce the concept of bias-variance trade-off.

Suppose we are interested in learning an unknown function $f(X)$ from the dataset $\mathcal{D}$, namely $\hat{f}(X;\mathcal{D})$. Then at some future query point $X = x_0$, we want to find $\hat{f}$, such that $\hat{f}$ satisfies:

$$\min_{\hat{f}} [f(x_0) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2$$

However are we really interested in evaluating this? Actually No! It is because the training dataset comes from a random selection! The prediction may not be accurate if we changed another training dataset.

In order to solve the problem, we should take the random selection process in an **average sense**.

Thus we are actually more interested in the following:

$$\min_{\hat{f}} \mathbb{E}_{\mathcal{D}}[f(x_0) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2$$

Theorem 1.2

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[f(x_0) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2 &= \underbrace{[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0]^2}_{\text{Bias}^2} \\ &+ \underbrace{\mathbb{E}_{\mathcal{D}}[[\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2]}_{\text{Variance}}\end{aligned}$$

Proof.

$$\begin{aligned}[f(x_0) - \hat{f}(x_0,\mathcal{D})|X = x_0]^2 &= [f(x_0) \underbrace{-\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) + \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))}_{=0} - \hat{f}(x_0,\mathcal{D})|X = x_0]^2 \\ &= [f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0]^2 + [\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]^2 \\ &+ 2[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0][\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]\end{aligned}$$

Now we take expectation to the expression above with respect to $\mathcal{D}$. We have:

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}[f(x_0) - \hat{f}(x_0,\mathcal{D})|X = x_0]^2 \\ = \ &\underbrace{[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0]^2}_{\text{Bias, as constant value}} \\ + \ &\underbrace{\mathbb{E}_{\mathcal{D}}\Big([\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]^2\Big)}_{\text{Variance}} \\ + \ &2\mathbb{E}_{\mathcal{D}}\Big\{[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0][\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]\Big\}\end{aligned}$$

Consider the last term, we expand this term, then we will have:

$$2\mathbb{E}_{\mathcal{D}}\Big\{[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0][\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]\Big\}$$

$$
\begin{aligned}
&= 2\mathbb{E}_{\mathcal{D}}\Big\{f(x_0)\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})] - f(x_0)\hat{f}(x_0;\mathcal{D}) - [\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]]^2 + \hat{f}(x_0;\mathcal{D})\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]\Big\} \\
&= 2\Big\{\mathbb{E}_{\mathcal{D}}\Big\{f(x_0)\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]\Big\} - \mathbb{E}_{\mathcal{D}}\Big\{f(x_0)\hat{f}(x_0;\mathcal{D})\Big\} - \mathbb{E}_{\mathcal{D}}\Big\{[\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]]^2\Big\} \\
&+ \mathbb{E}_{\mathcal{D}}\Big\{\hat{f}(x_0;\mathcal{D})\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]\Big\}\Big\} \\
&= \mathbb{E}_{\mathcal{D}}(f(x_0))\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \mathbb{E}_{\mathcal{D}}(f(x_0))\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))^2 + \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))^2 \\
&= 0
\end{aligned}
$$

The last term vanished, since we know that $\mathbb{E}_{\mathcal{D}}(f(x_0)) = f(x_0)$ and $\mathbb{E}_{\mathcal{D}}(\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))) = \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))$.

Also do note that even though $\hat{f}(X;\mathcal{D})$ is a random variable due to the randomness of the dataset, $\mathbb{E}_{\mathcal{D}}\hat{f}(X;\mathcal{D})$ is no longer a random variable with respect to $\mathcal{D}$.

Example 1.3
If $Z \sim N(\mu, \sigma^2)$ is a random variable, then $\mathbb{E}(Z) = \mu$ is no longer random variable.

Note 2
The reason for us to condition on x_0, is because we want to simplify the entire process.
In fact, by using total expectation law, we are able to evaluate for $\mathbb{E}_{\mathcal{D}}\big[f(X) - \hat{f}(X;\mathcal{D})\big]^2$

Theorem 1.3 (Total Expectation Law)
Given any random variables X, Y, then we have

$$\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X|Y)\Big)$$

Thus we have:

$$\mathbb{E}_{\mathcal{D}}\big[f(X) - \hat{f}(X;\mathcal{D})\big]^2 = \mathbb{E}_X\Big[\mathbb{E}_{\mathcal{D}}\big[f(X) - \hat{f}(X;\mathcal{D})\big]^2 | X = x_0\Big]$$

Note 3
Practically, $\mathbb{E}_{\mathcal{D}}\Big(\hat{f}(x_0;\mathcal{D})\Big)$ can be estimated by:

$$\mathbb{E}_{\mathcal{D}}\Big(\hat{f}(x_0;\mathcal{D})\Big) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(b)}(x_0)$$

From this, we can estimate the variance term by:

$$\mathbb{E}_{\mathcal{D}}\Big[[\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2\Big] = \frac{1}{B}\left(\sum_{b=1}^{B}\left[\hat{f}(x_0) - \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(b)}(x_0)\right]\right)$$

This implies that, to estimate for variance, it is not necessary for us to know the ground-truth model. However, to know the bias, we do need to know the ground-truth model.

Definition 1.3 (Bias and Variance)
Given an estimator $\hat{\theta}$ for some parameter θ, the bias is defined as:

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

while given a random variable X, the variance is defined as:

$$\text{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}(X))^2\big]$$

Suppose we are interested in a model that perfectly fits into the data. For example,

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

This is the population-level model we seen before. Let $\text{Var}(\varepsilon) = \sigma^2$ and $\mathbb{E}(\varepsilon) = 0$,

We now take samples from such model, we will then have (x_i, y_i) that is in relationship of:

$$y_i = f(x_i) + \varepsilon_i \qquad i = 1, ..., n$$

We are interested in finding g, where g can minimizes:

$$\sum_i (y_i - g(x_i))^2$$

After solving the optimization problem, we will obtain $\hat{g}$, that $\hat{g}(x_i) = y_i, i = 1, ..., n$.

To adopt bias-variance view, we will have to evaluate $\mathbb{E}(\hat{g})$ and $\text{Var}(\hat{g})$.

Note that:

$$\begin{aligned} \mathbb{E}(\hat{g}(x_i)) &= \mathbb{E}(y_i) \\ &= \mathbb{E}(f(x_i) + \varepsilon_i) \\ &= \mathbb{E}(f(x_i)) \\ &= f(x_i) \end{aligned}$$

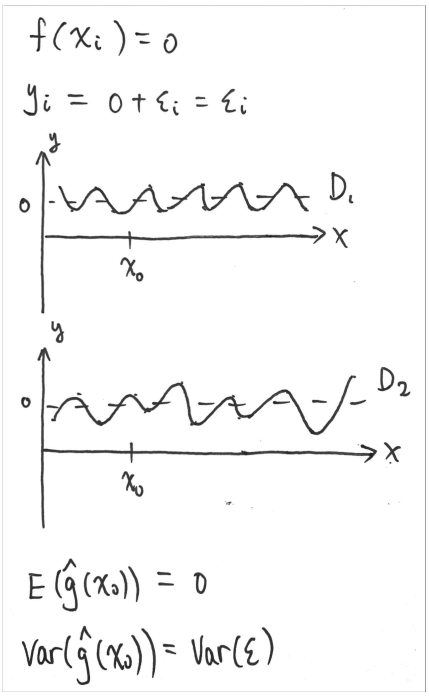
To see if our estimator having bias or not, we only need to check $\mathbb{E}(\hat{g}(x_i)) - f(x_i)$.

However, since $\mathbb{E}(\hat{g}(x_i)) = f(x_i)$, thus $\mathbb{E}(\hat{g}(x_i)) - f(x_i) = 0$. This implies if the model is very flexible that can perfectly fits into the data, then the model is unbiased.

Now, consider the variance, we have:

$$\begin{aligned} \text{Var}(\hat{g}(x_i)) &= \text{Var}(y_i) \\ &= \text{Var}(f(x_i) + \varepsilon_i) \\ &= \text{Var}(\varepsilon_i) \\ &= \sigma^2 \end{aligned}$$

For a very flexible model, the bias can be very small, while the variance will be extremely large.



Suppose there is an dataset $\mathcal{D}$, and we want to solve the following optimization problem:

$$\min_g \sum_{i=1}^{n} \big((y_i - g(x_i))\big)^2 + \lambda \cdot R(g)$$

If $g(x_i) = C$ is a constant function (That is, totally not fitting into the data).

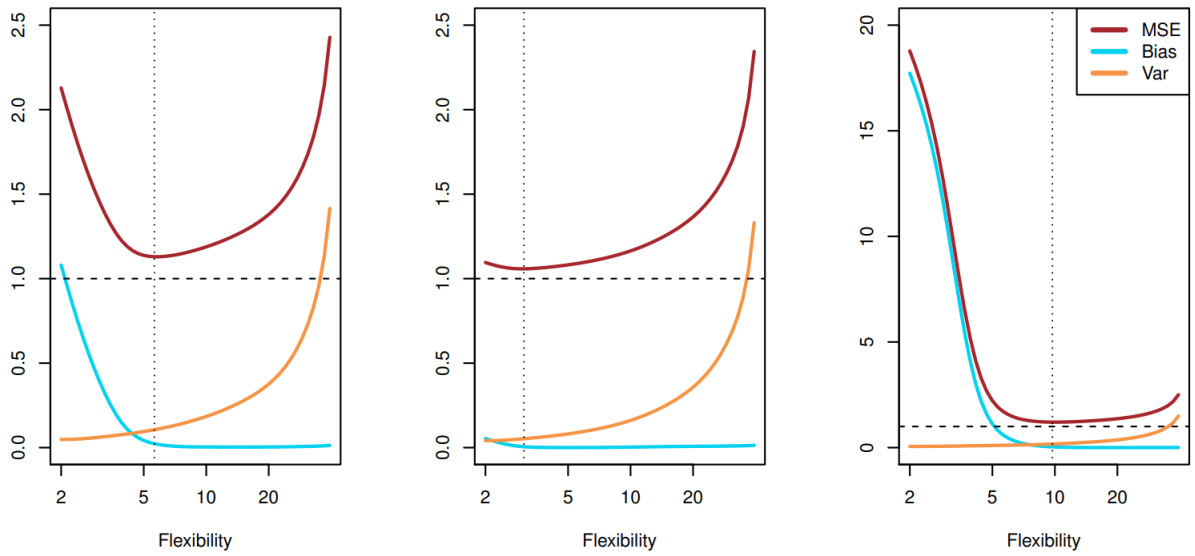
Since $\mathbb{E}(\hat{g}(x_i)) = C$, and $\mathrm{Var}(\hat{g}) = 0$.

Thus

$$\begin{aligned} \text{Bias} &= \mathbb{E}(\hat{g}(x_0)) - f(x_0) \\ &= C - f(x_0) \end{aligned}$$

Which is not reducible. This implies the model is biased. However, the variance is 0.

To conclude, we want to find a model, which can keeps a perfect balance between bias and variance. The model may be good, even if the model is biased.



Red line: Test MSE	Blue line: Bias	Orange line: Variance

2 Regression analysis

2.1 Simple Linear Regression

Definition 2.1 (Simple Linear Regression)
In simple linear regression, there are only 1 variable, and thus there is only 2 parameters to fit in, which is:

$$Y = \underbrace{\beta_0 + \beta_1 X}_{f(X)} + \varepsilon$$

Suppose we want to estimate $\hat{\beta}_0, \hat{\beta}_1$, then we have the realization value to be:

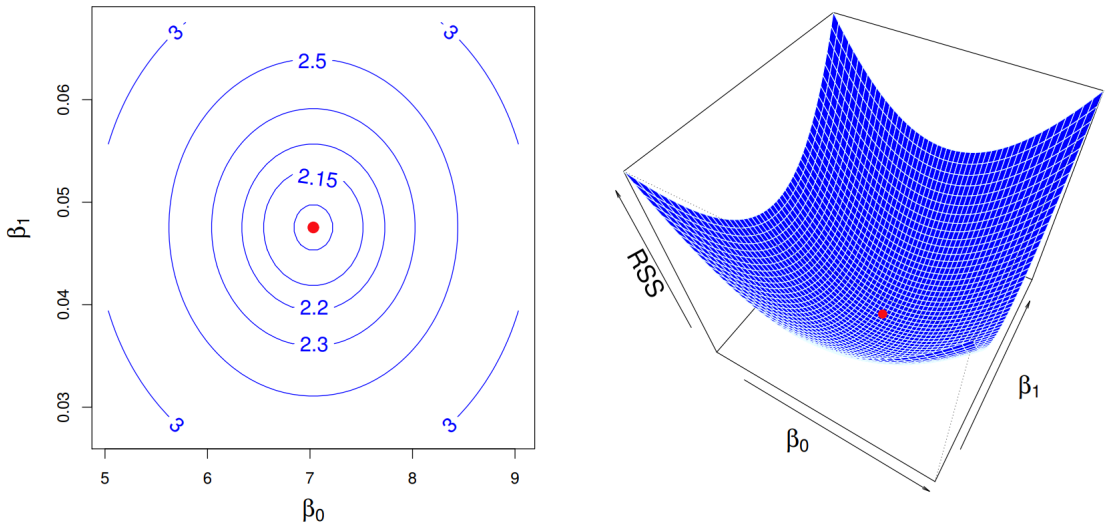
$$\hat{Y} = \beta_0 + \beta_1 X$$

To figure out the estimate, suppose there is some observed training data $D = \left\{(x_i, y_i)\right\}_{i=1\ldots n}$.

What we want to do now is to solve the optimization problem on the squared loss

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

This problem is also known as least square problem, and the figure below shows the function that we want to find the optimized value.



Theorem 2.1 (Least Square Problem)
The solution to the least square problem is given by:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \hat{\beta}_0 &= \bar{y}-\hat{\beta}_1\bar{x}\end{aligned}$$

Question
Given $\mathcal{D}^{(1)}$, we can generate $\hat{\beta}_0{}^{(1)}$ and $\hat{\beta}_1{}^{(1)}$.
Given $\mathcal{D}^{(2)}$, we can generate $\hat{\beta}_0{}^{(2)}$ and $\hat{\beta}_1{}^{(2)}$. etc.
In an average sense, will $\hat{\beta}_0{}^{(1)} = \beta_0^*$? i.e. does

$$\mathbb{E}_{\mathcal{D}}\left(\beta_0{}^{(1)}\right) = \beta_0^*$$

Answer
Unbiased

Consider the variance of $\hat{\beta}_0$, which is $\text{Var}(\hat{\beta}_0)$. We always wanted to keep the variance as small as possible. The following shows an example of the estimation of $\hat{\beta}_0$. Note that both of them are unbiased, however, the one on the right hand side obviously have a larger variance than the left hand side.

To be continued... [2024-09-19, 00:00]