Machine Learning

Samatrix Consulting Pvt Ltd

What is Machine Learning?

Machine Learning

- To start the introduction to Machine Learning, let’s start with a simple example.
- Suppose you have been assigned as Data Scientist to advice on how to improve the sales of a particular product of a company.
- The company provided you with sales data from 200 different markets.
- The data also contains the advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.
- The client cannot directly increase the sales of the product.
- But they can adjust the advertisement budget for each of the three media.

Machine Learning

- As a data scientist, if you can establish the relationship between advertisement expenditure and sales, you can provide your feedback on how to adjust the budgets so that sales can increase.
- So, the objective is to develop a model that you can use to predict the sales on the basis of the three media budgets.

Machine Learning

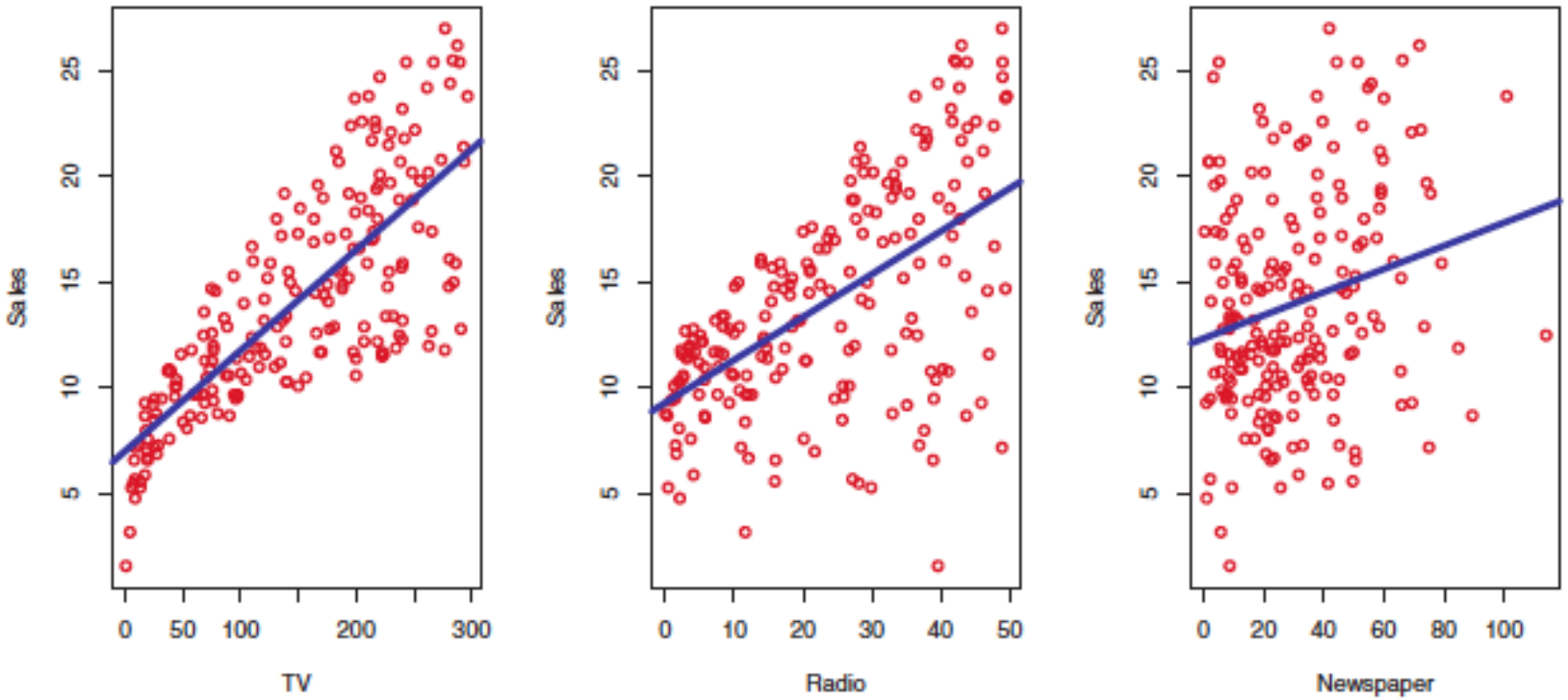


Fig 1: The plots display sales in thousands of units. Sales is a function of budget on TV, Radio, and Newspaper in thousands of dollars across 200 markets

Input – Output Variables

- In this case, the three media budgets are input variables.
- We generally denote input variables by symbol X and use subscripts to distinguish among the input variables.
- Hence, we can denote TV budget by X_1, Radio budget by X_2, and newspaper budget by X_3.
- In this case, the predicted sales is the response of the model.
- We call the response by output variable and denote the response by symbol Y.

Independent Vs Dependent Variables

- We can give the input variables different names such as independent variables, predictors, features etc.
- The output variable or response is also known as the dependent variable.
- If we observe one quantitative output variable Y and p different input variables, $X_1, X_2, \ldots, X_p$.
- Our assumption is that there is a relationship between Y and $X = (X_1, X_2, \ldots, X_3)$. We can write the relationship as

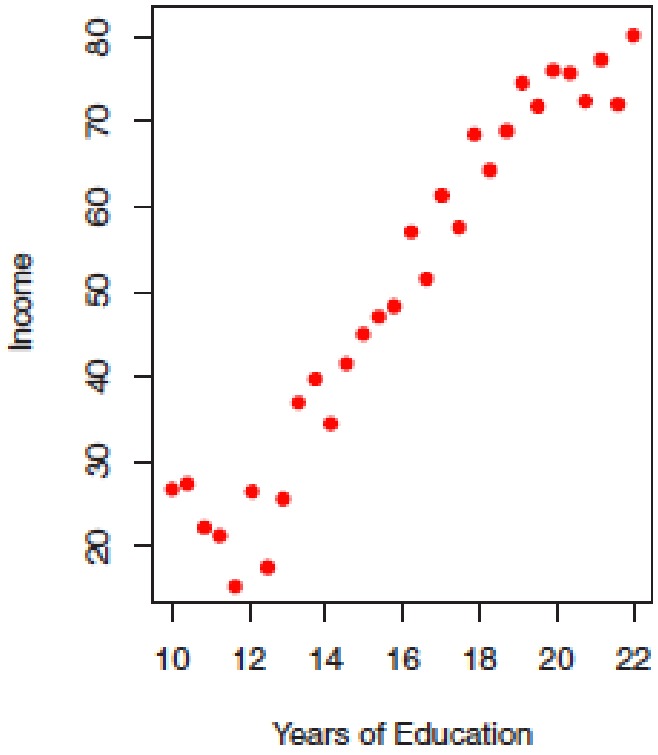
$$Y = f(X) + \epsilon$$

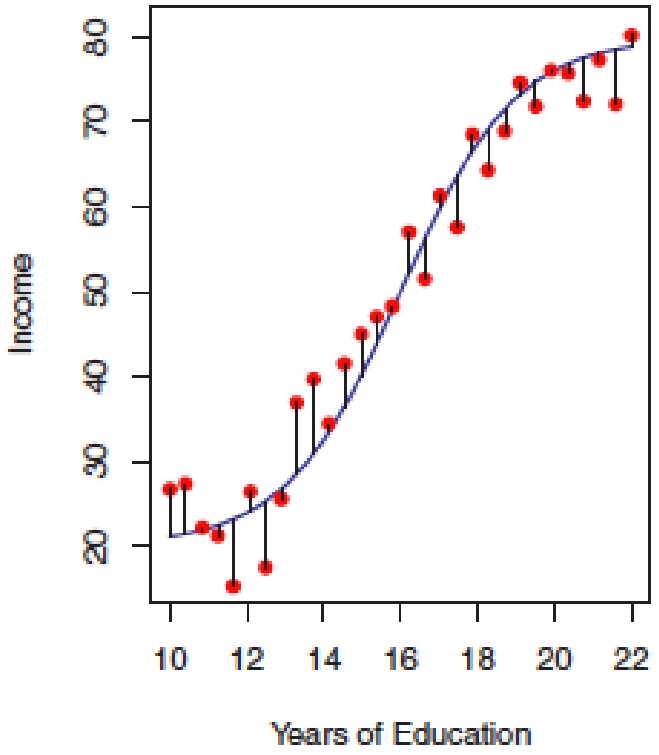
- In this case f is some fixed but unknown function of $X_1, \ldots, X_p$. ϵ is random error that is independent of X. The mean value of ϵ is zero.

Function Approximation

Function Approximation

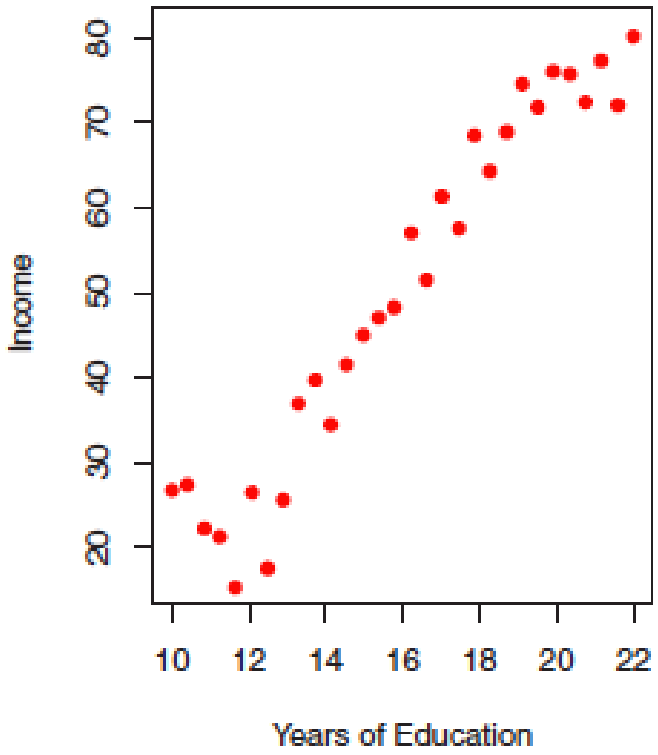
- Let us consider another dataset, income dataset.
- The left-hand panel shows the plot of income versus years of education for 30 individuals.
- Using the plot, you may be able to predict the income given the years of education.
- But the function that relates the income to the years of education is not known.

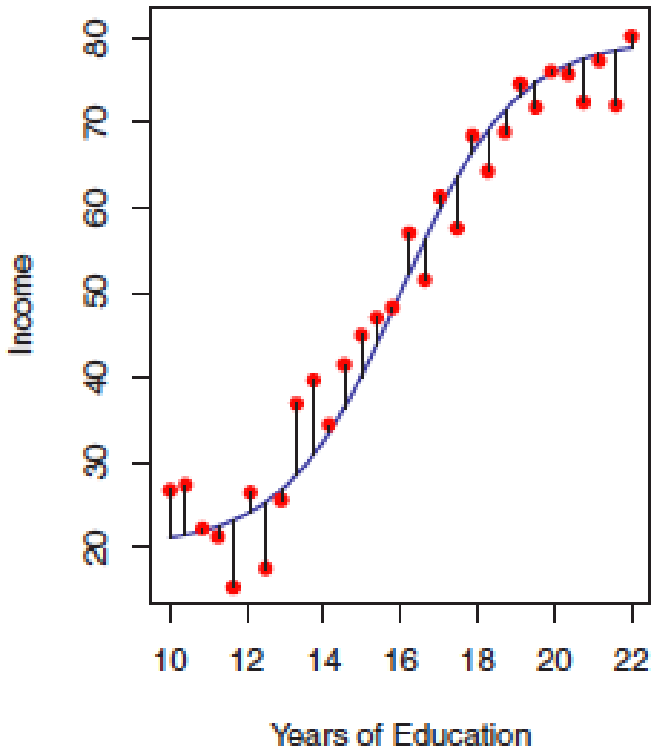




Function Approximation

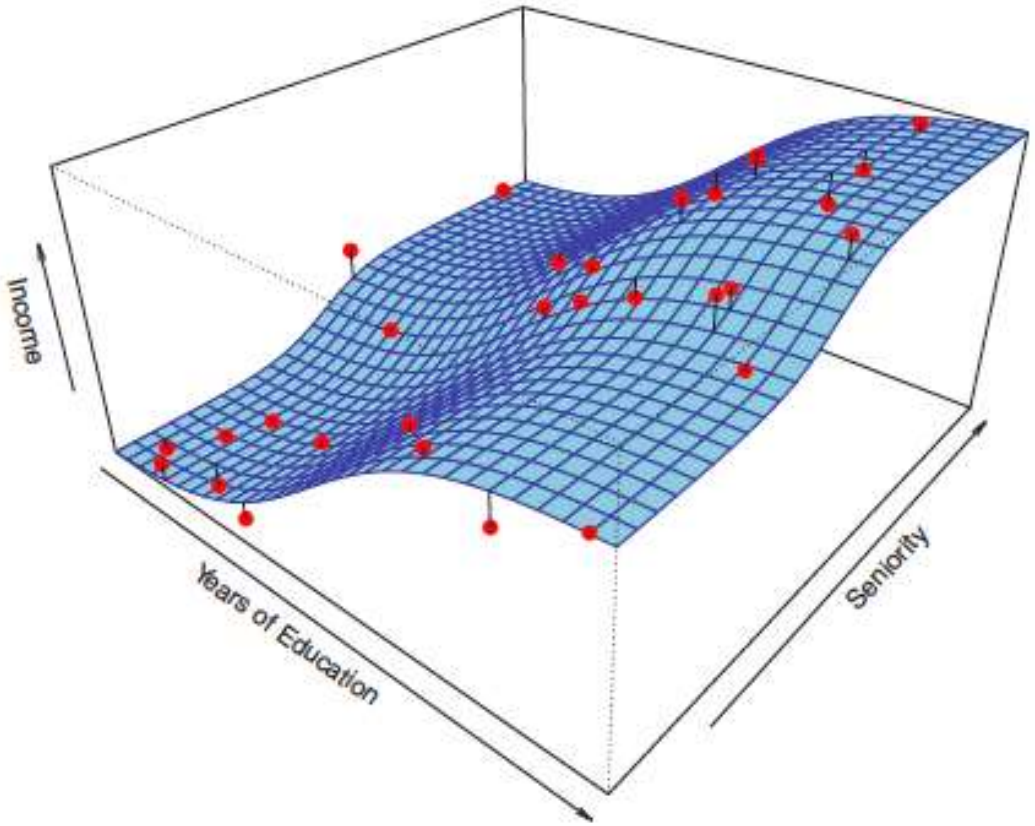
- In this situation, we should estimate the function f based on observed data.
- Since the dataset is simulated, so we know the function f.
- Based on the given function, we have plotted the blue curve in the right-hand side panel.
- You may notice the vertical lines that represent the error terms ϵ.
- Out of the 30 observations, some observations lie above the blue curve whereas some observations lie below the blue curve.
- But the overall mean would be zero.





Function Approximation

- Generally, the function f may involve more than one input variable.
- In the figure, the plot of income as a function of seniority and years of education is given.
- We need to estimate the function f that is two-dimensional surface.
- So, we can say that machine learning refers to a set of approaches for estimating function f.



Why to estimate f?

Why Approximate f

- The purpose of function approximation is either prediction or inference.

Prediction

- In several cases, a set of input variable X are available but we cannot easily obtain the response variable Y. Since the average of error terms is zero, we can predict Y using
- $\hat{Y} = \hat{f}(X)$
- In this case, $\hat{f}$ represents the estimated f and $\hat{Y}$ represents the predicted outcome of Y.If the prediction is the priority, the $\hat{f}$ is often treated as black box if it can provide accurate estimation of Y.

Prediction

- Example: Suppose $X_1, \ldots, X_p$ represent the characteristics of a patient's blood sample that has been measured in a lab.
- The output variable Y encodes the patient's risk for a severe adverse reaction to a particular drug.
- We can predict Y using X so that we can avoid giving the drug to the patient who are at high risk of an adverse reaction.
- That means we can predict the set of patients who are at high risk of an adverse reaction.

Reducible - Irreducible Errors

- Two quantities, reducible error and irreducible error, define the accuracy of $\hat{Y}$ as a prediction of Y.
- The $\hat{f}$ cannot be the perfect estimator of f.
- Due to this some errors will be introduced.
- We can reduce such error by improving the accuracy of $\hat{f}$ by using the most appropriate machine learning technique to estimate f.
- Even if, we could perfectly estimate f, our estimated response would be $\hat{Y} = f(X)$.
- The prediction would still have some error because Y is also a function of ϵ and we cannot predict ϵ using X.
- The variability associated with ϵ affects the accuracy of our predictions.
- This is known as **the irreducible error**. Irrespective of our accuracy in estimating f, some errors that have been introduced by ϵ, cannot be reduced.

Reducible - Irreducible Errors

- Why the irreducible errors are larger than zero?
- The quantity ϵ may contain some unmeasured variable that can be used to predict Y.
- Because we do not measure such variables, f cannot used them for prediction.
- Unmeasurable variations are also included in the quantity ϵ.
- For example, due to manufacturing variation in the drug or the general feeling of well-being of the patient on the given day.

Reducible - Irreducible Errors

- If the estimated function is $\hat{f}$ and set of predictors is X. We get the prediction $\hat{Y} = \hat{f}(X)$. We can show that

$$
\begin{aligned}
E\left(Y - \hat{Y}\right)^2 &= E\left[f(X) + \epsilon - \hat{f}(X)\right]^2 \\
&= E\left[f(X) - \hat{f}(X)\right]^2 + Var(\epsilon) \\
&= Reducibale + Irreducible
\end{aligned}
$$

- Where $E\left(Y - \hat{Y}\right)^2$ represents the average, or expected value, of the squared difference between the predicted and actual value of Y and $Var(\epsilon)$ represents the variance associated with the error term ϵ.

Inference

- On various occasions, we want to understand how Y changes as the we change $X_1, \ldots, X_p$.
- We wish to estimate f but making predictions for f may not be our goal.
- We may want to understand the relationship between X and Y.
- In such situations, we need to know understand the exact form of $\hat{f}$. Hence, it cannot be treated as a black box.
- We would like to answer the following questions:

Inference

- **Which input variables are associated with output variable Y?**
- Generally, a small fraction of the available input variables is associated with response Y.
- It is extremely useful to understand few important predictors out of a big set of all the possible predictors.

Inference

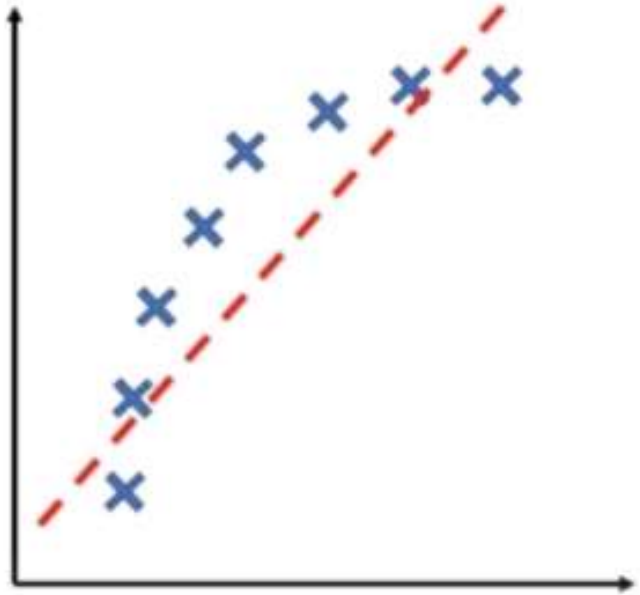
- **Understand the relationship between the response and each predictor?**
- Different predictors may have different relationship with the response.
- Some predictors may have a positive relation with Y whereas some other predictors may have a negative relationship with Y.
- If the relationship is positive, the increase in the value of predictor results in the increase in the value of Y.
- Negative relationship has opposite relationship.
- Due to the complexity of the function f, the relationship between a predictor and the response may also depend upon the values of the other predictors.

Inference

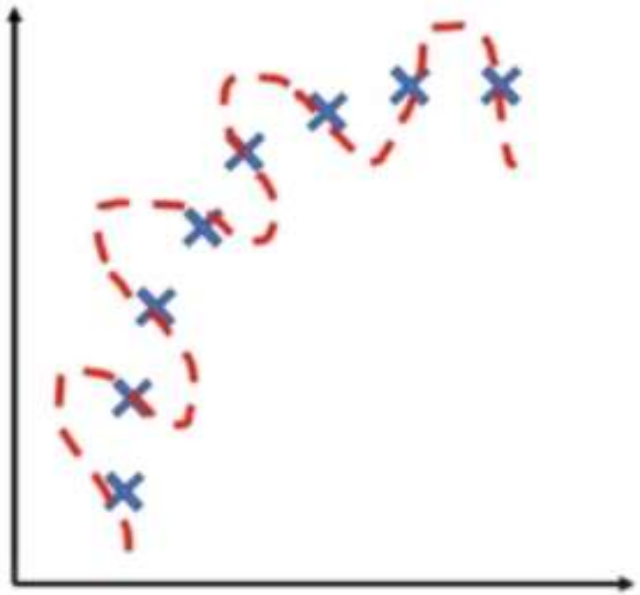
- **Whether the relationship between Y and each predictor is linear or more complex?**
- Generally, the methods for estimating f are linear.
- However, in certain cases, the relationship is more complex that cannot be accurately represented by a linear model.

How to estimate f ?

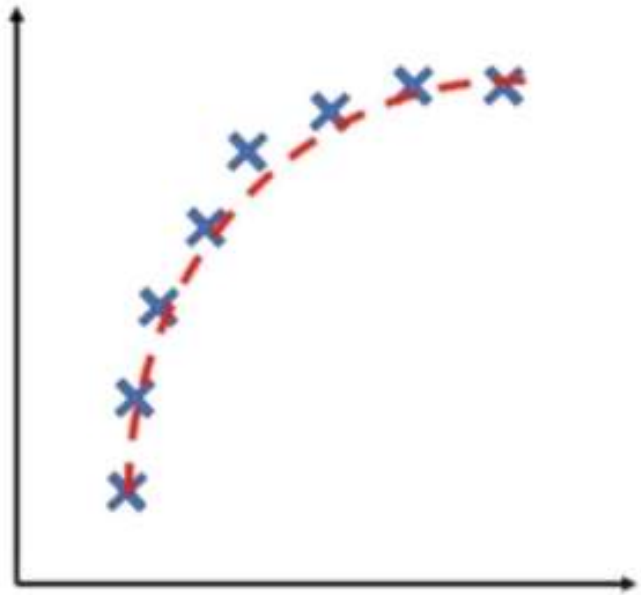
Underfitting



Overfitting

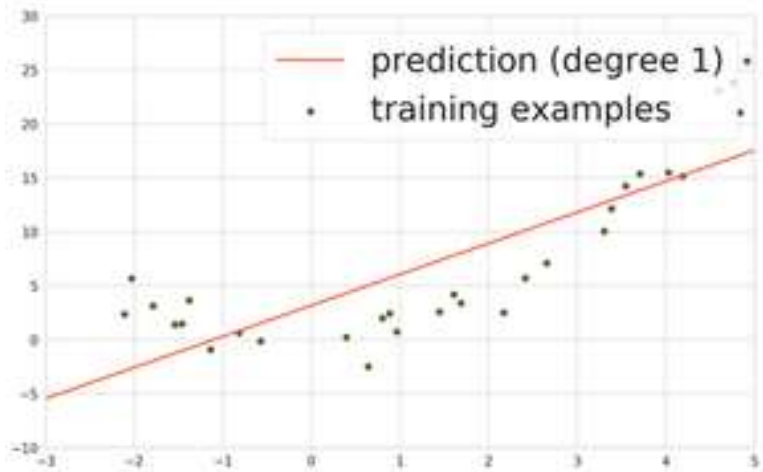


Ideal Balance



Underfit

High bias

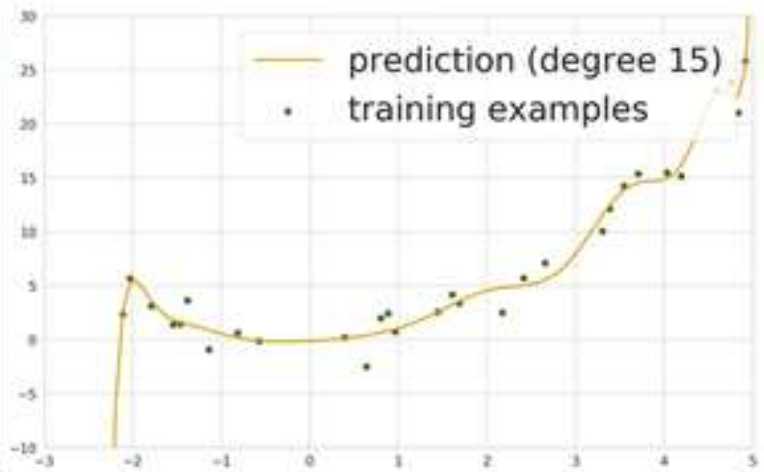


Good Fit

Low bias, low variance

Overfit

High variance



Types of Model Fit

How to estimate f

- Suppose we have a set of observed data that includes $n = 30$ data points.
- We name these observations, training data. The training data will be used to train or teach our model to estimate f.
- Let x_{ij} represents the value of j^{th} predictor or input for observation i where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$.
- Correspondingly, the response variable for the i^{th} observation is represented by y_i.
- Hence our training data consists of $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$.

How to estimate f

- Our objective is to estimate the unknown function f.
- For this, we apply the machine learning model to the training data.
- We want to find the function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y).
- We can use two approaches for the same: parametric and non-parametric approach.

Parametric model	Non-parametric model
It uses a fixed number of parameters to build the model.	It uses flexible number of parameters to build the model.
Considers strong assumptions about the data.	Considers fewer assumptions about the data.
Computationally faster	Computationally slower
Require lesser data	Require more data
Example – Logistic Regression & Naïve Bayes models	Example – KNN & Decision Tree models

Parametric Approach

- Parametric approach is 2 step model-based approach
- First step is to make an assumption about the functional form or shape of f. For example, we can make a simple assumption that f is linear in X:

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

- Once the assumption about a linear f is made, the f can be estimated easily.
- Instead of estimating the complete p-functional function $f(X)$, we have to estimate $p + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$

Parametric Approach

- After we select the model, we use the training data to fit or train the model.
- From the example given in step 1, we need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$.
- In other words, we need to find the values of the parameters such that

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Parametric Approach

- Several ways are available to fit the linear model. (Ordinary) least square method is one of the possible ways to fit the linear model.
- Least square method is the most common approach.
- The model-based approach mentioned above is known as parametric approach.
- This approach reduces the problem of estimating f to estimate a set of parameters.
- Using the linear model, on one hand it is very easy to estimate the parameters, such as $\beta_0, \beta_1, \ldots, \beta_p$, on the other hand, the model that we choose may not represent the true unknown form of f.

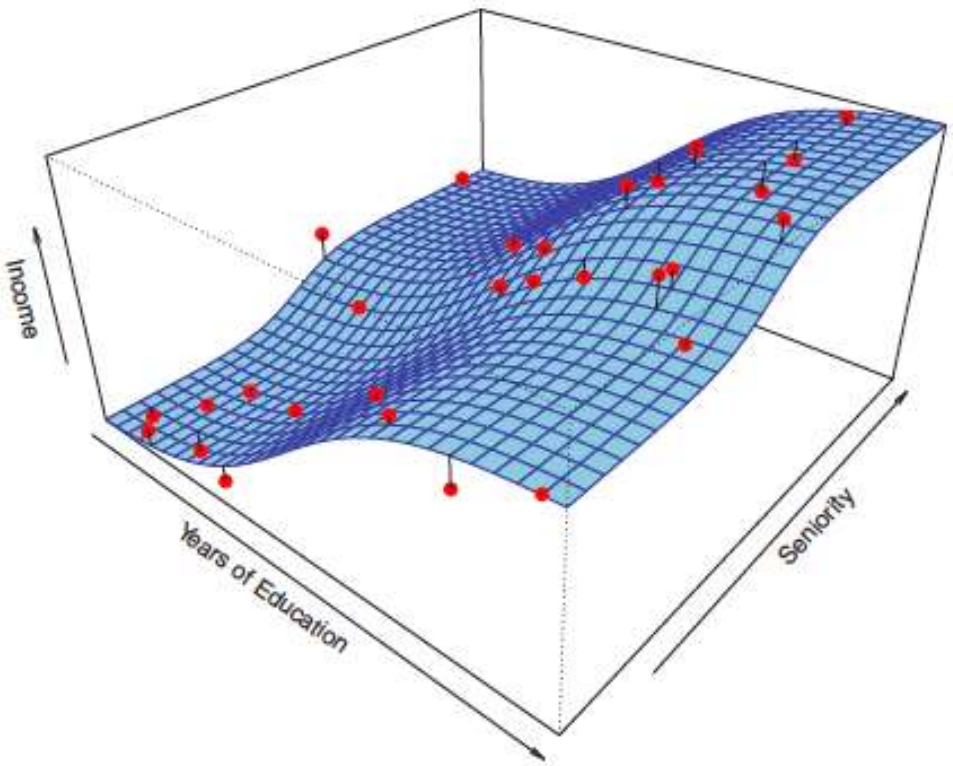
Parametric Approach

- If the estimated model is too far from the true f, our estimate will not be good.
- To fix this problem, we can choose the flexible model but such models require estimating a greater number of parameters.
- The more complex models may lead to the overfitting of the data which means that the models follow the errors, or noise, too closely.
- We applied the parametric approach to the Income data as shown in the figure below. In this case we fit the linear model to the form

$$income \approx \beta_0 + \beta_1 \times education + \beta_2 \times seniority$$

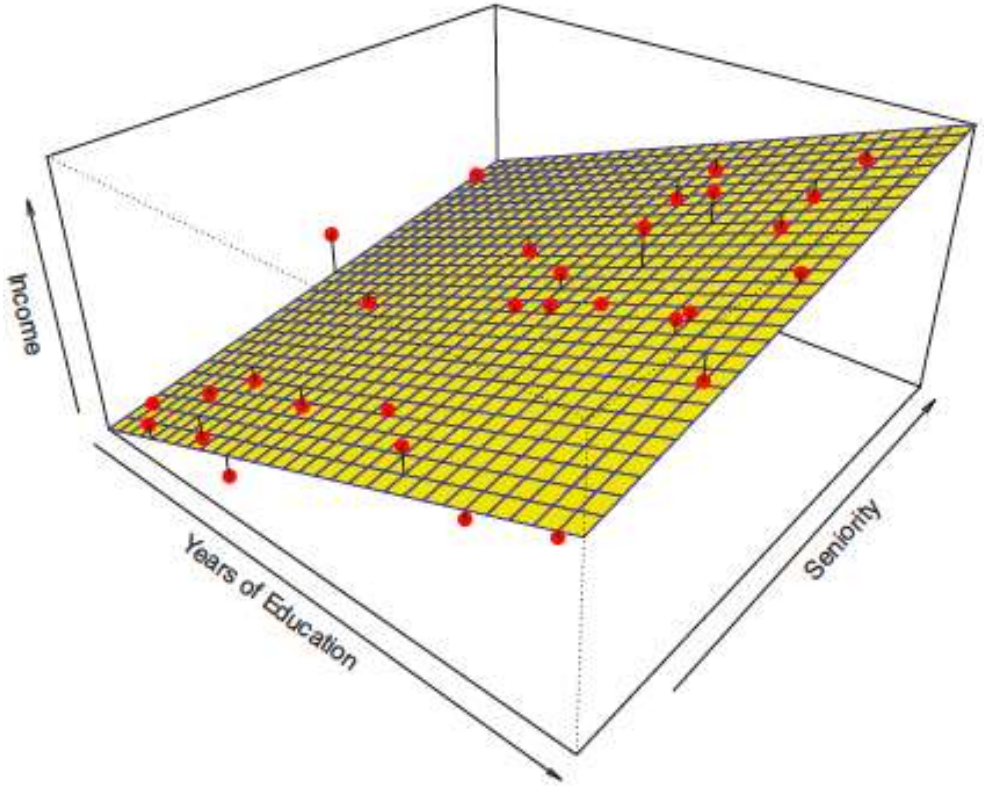
Parametric Approach

- In this case, we have assumed the linear relationship between the response and the two predictors.
- Hence, we have to estimate the values of $\beta_0, \beta_1, and\ \beta_2$ that we can estimate using the least squares linear regression.
- The following picture shows the plot of income as a function of years of education and seniority in the Income data set.
- The blue surface represents the true underlying relationship between income and years of education and seniority



Parametric Approach

- The picture below a linear model fit by least squares to the Income data from previous picture.
- The observations are shown in red, and the yellow plane indicates the least squares fit to the data.
- When we compare both the figures, we can see that the linear fit in the second figure is not correct.
- The true f has some curvature that the linear fit does not capture.
- But the linear file has done a reasonable job of capturing the positive relationship between years of education and income, as well as the slightly less positive relationship between seniority and income.



Non - parametric Approach

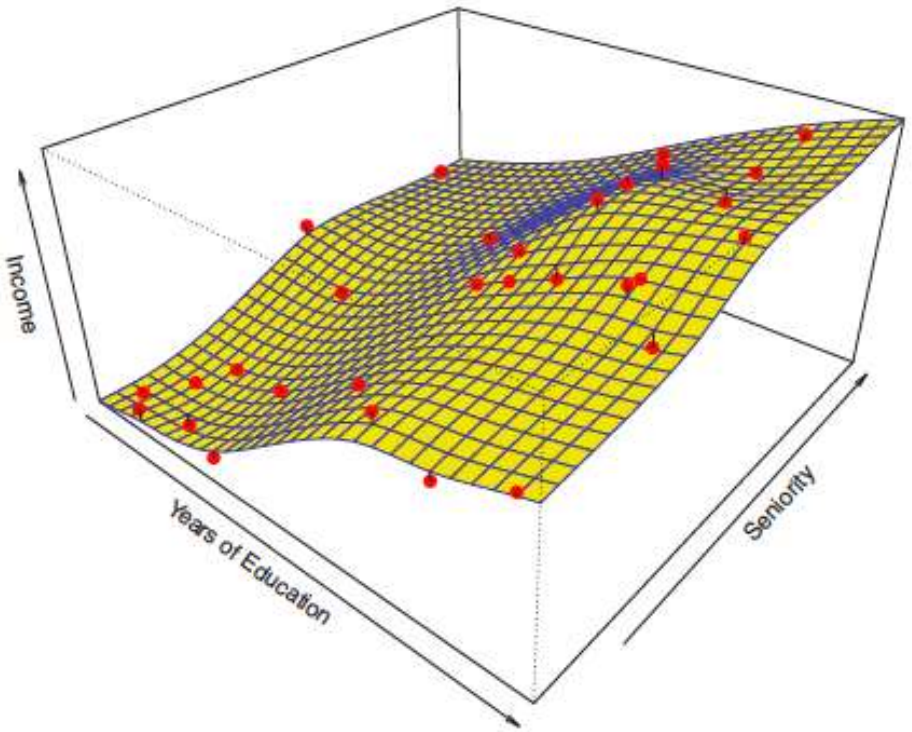
- With non-parametric methods, we do not make explicit assumptions about the functional form of f.
- These methods estimate f that is as close to the data points as possible without being too rough or wiggly.
- Non-parametric methods have an advantage over the parametric methods.
- Non-parametric methods avoid the assumption of a particular functional form for f.
- Hence, they can accurately fit a wider range of possible shapes for f.

Non - parametric Approach

- Some risks are involved with the parametric approach.
- The functional form that we used to estimate f may be very different from the true f.
- In this case, the resulting model will not fit the data well.
- On the other hand, non-parametric methods completely avoid this risk.
- Be we do not make any assumption about the form of f.
- However non-parametric methods have a disadvantage.
- These approaches do not reduce the number of parameters to estimate f, we need a larger number of observations when compared to parametric methods to estimate f.

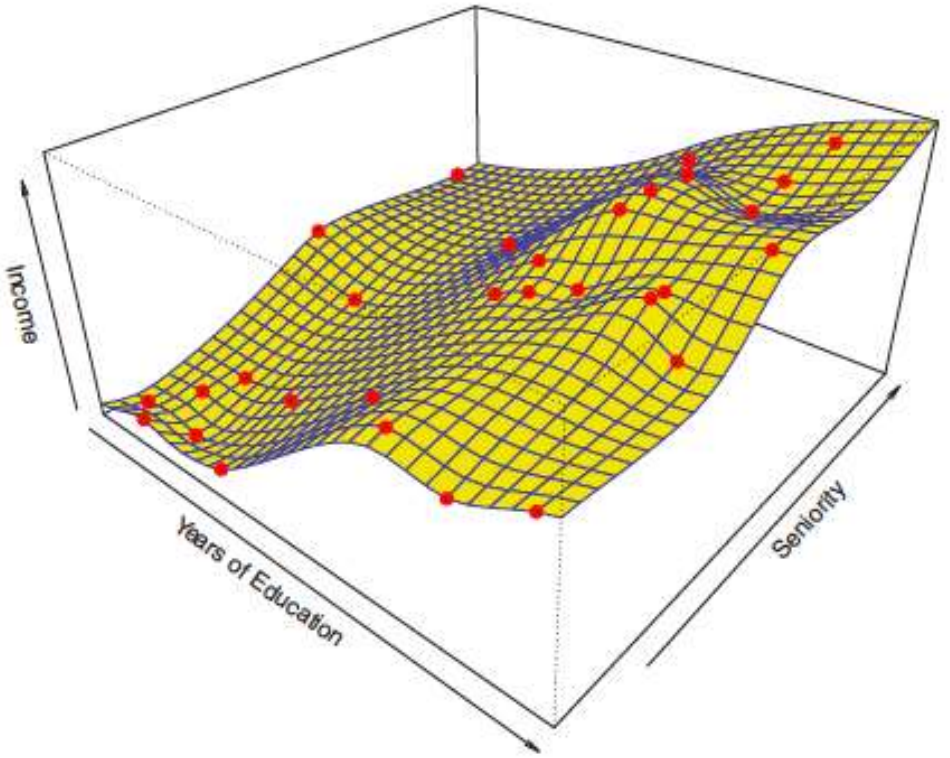
Non - parametric Approach

- The figure below shows the results of a non-parametric approach (thin-plate spine) to fit the Income data.
- No pre-specified model was not imposed on f.
- It tries spline to produce an estimate for f.
- This is as close as possible to the observed data. In the picture below, it is shown by yellow surface.
- It is a smooth thin plate spline fit. In this case, the non-parametric fit has resulted in a reasonably accurate fit estimate of true f.



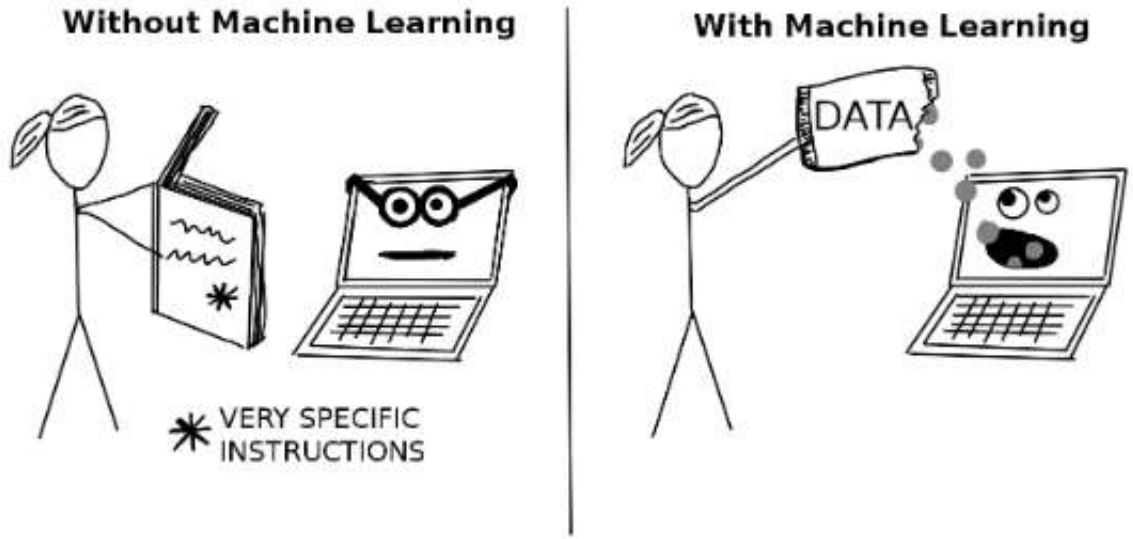
Non - parametric Approach

- For the thin-plate spline, the data analyst should select a level of smoothness that allows a rougher fit.
- With the increase in roughness, the resulting estimate fits the observed data perfectly so that the errors are zero.
- But the rough spline fit is far more variable than the true function f.
- This is an example of overfitting the data.
- It is an undesirable situation because the fit obtained will not yield accurate estimates of the response on new observations that were not part of the original training data set.

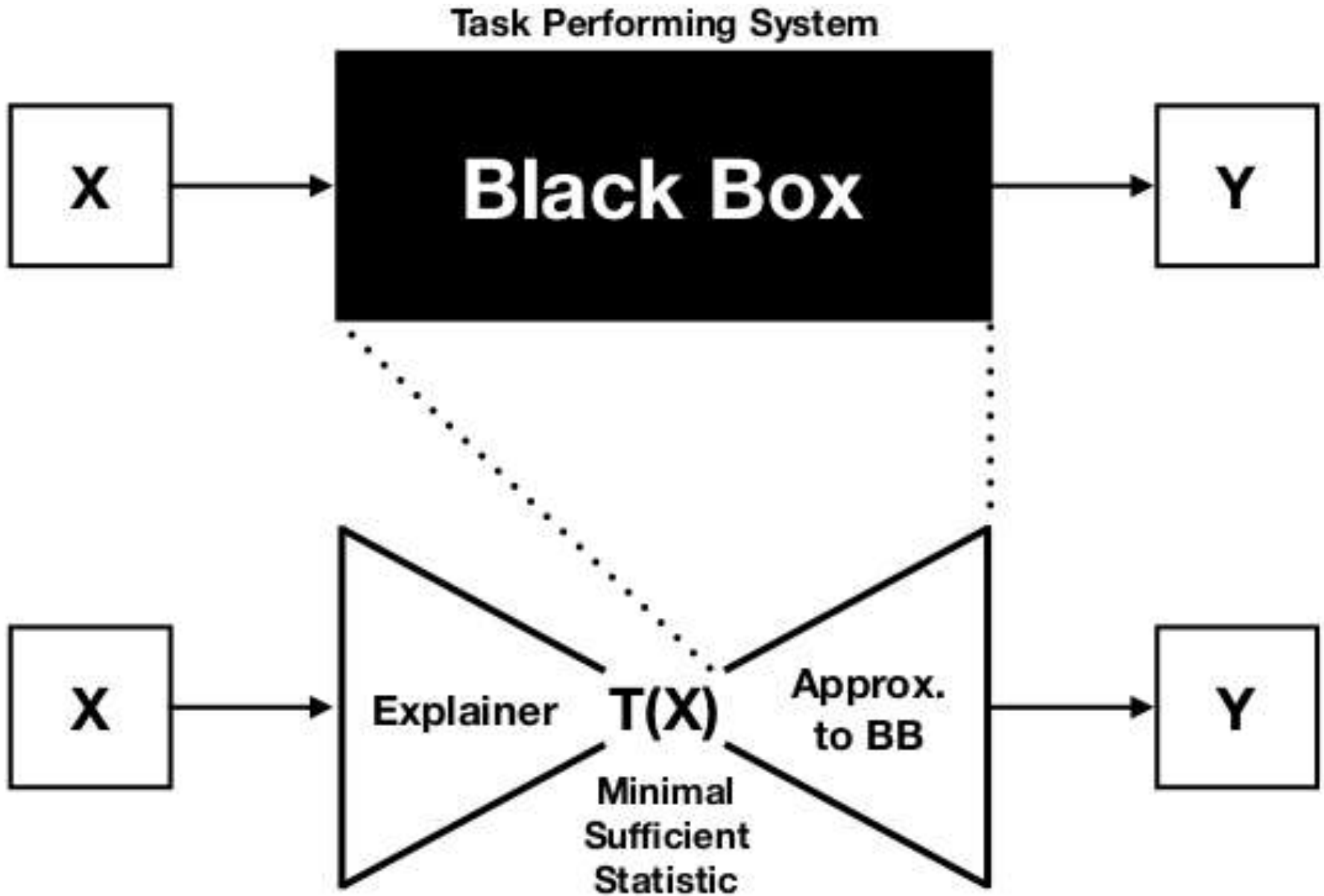


ADVANTAGES	DISADVANTAGES
Non-Parametric algorithms **do not make** any **assumptions** on functional forms & work on trying to best fit the data.	But, require **large amounts of training data** for estimating & constructing mapping functions.
Capable of **fitting & learning** a **large number of functional forms** (since no constraints of just one type of form)	They have many parameters to train & work on. Hence, they are **SLOWER** in giving results
They result in **high performance models** for prediction.	Large amounts of training data may result in **OVERFITTING** & **no justifications** available for certain **predictions** made.

Prediction Accuracy vs Model Interpretability

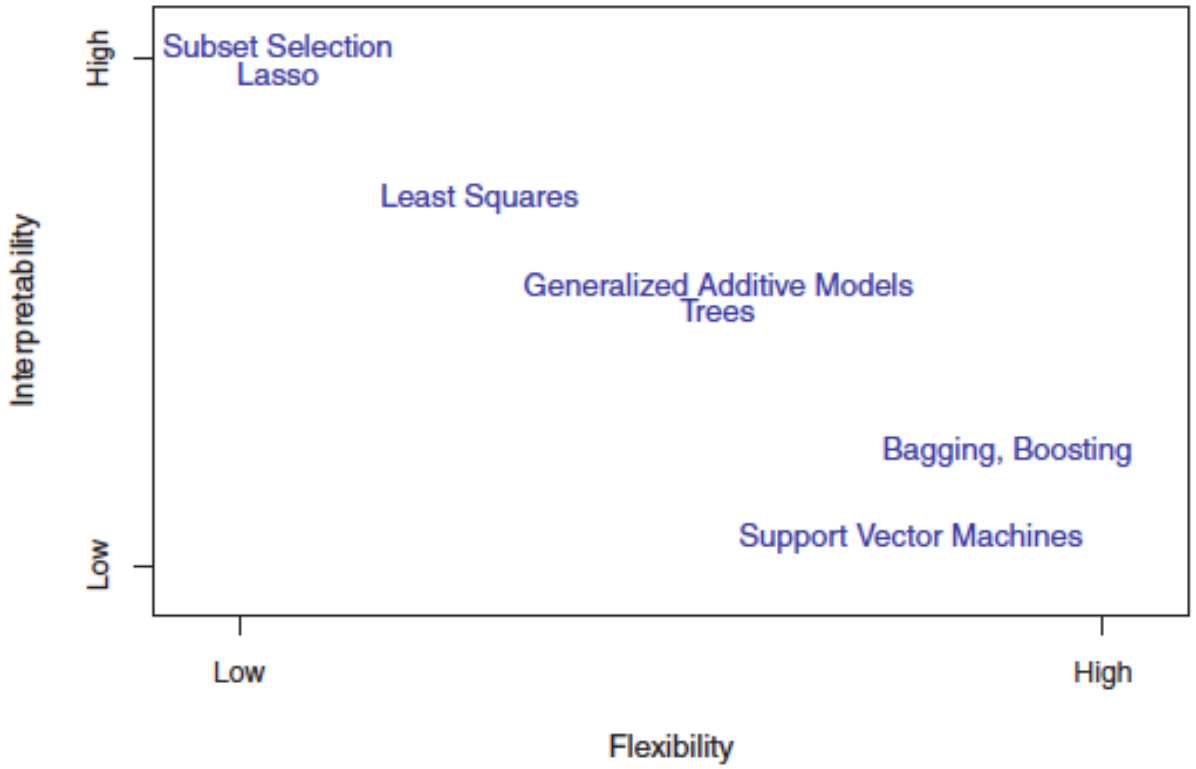

Without Machine Learning
With Machine Learning
DATA
*VERY SPECIFIC INSTRUCTIONS


Why did you predict 42 for this data point?
awkward silence


Task Performing System
X
Black Box
Y
X
Explainer
T(X)
Minimal
Sufficient
Statistic
Approx.
to BB
Y

Trade Off

- Out of several machine learning methods, some methods are less flexible.
- That means that they are more restrictive and they can produce just a relatively small range of shapes to estimate f.
- For example, linear regression is a relatively inflexible approach, because it can only generate linear functions.
- On the other hand, methods such as the thin plate splines are considerably more flexible because they can generate a much wider range of possible shapes to estimate f.



Trade Off

- We can ask the following question:
- why should we choose to use a more restrictive approach instead of a very flexible method?
- Due to several reasons, we might prefer a more restrictive model. If our goal is inference, then restrictive models are more interpretable.
- For example, if we are interested in inference, we may choose a linear model, because the linear model helps us understand the relationship between Y and $X_1, X_2, \ldots, X_p$ easily.
- On the other hand, very flexible approaches, such as the splines and the boosting methods can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.

Trade Off

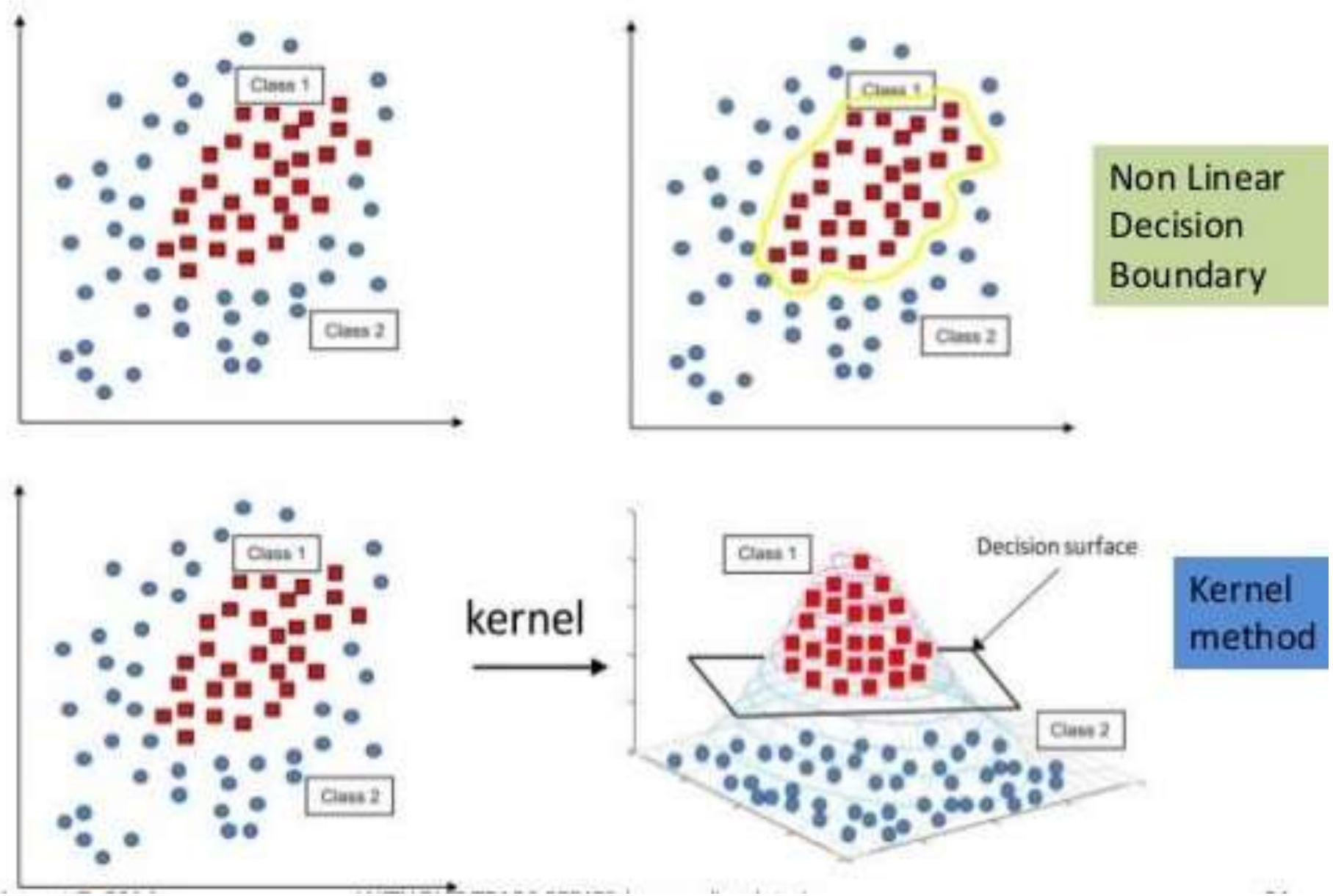
- The figure above underlines the trade-off between flexibility and interpretability for some of the machine learning methods.
- Least squares linear regression is relatively inflexible but is quite interpretable.
- We can take the case of the lasso method.
- This method relies upon the linear model but it uses an alternative fitting procedure for estimating the coefficients $\beta_0, \beta_1, \dots, \beta_p$.
- The lesso method is more restrictive in estimating the coefficients, and sets a number of them to exactly zero.
- Hence, we can state that the lasso is a less flexible approach than linear regression.
- It is also more interpretable than linear regression, because in the final model the response variable will only be related to a small subset of the predictors—namely, those with nonzero coefficient estimates.

Trade Off

- On the other hand, generalized additive models (GAMs) extend the linear model to allow for certain non-linear relationships.
- Hence, GAMs are more flexible than linear regression.
- However, they are less interpretable than linear regression, because the relationship between each predictor and the response is now modeled using a curve.
- Finally, fully non-linear methods such as bagging, boosting, and support vector machines with non-linear kernels are highly flexible approaches that are harder to interpret.

Trade Off

- Hence, we can state that when inference is the goal, we should use simple and relatively inflexible machine learning methods.
- However, there might be situations, when we are only interested in prediction, and not in the interpretability of the predictive model.
- For instance, if we want to build a model to predict the price of a stock, we would be interested in an algorithm that can predict accurately whereas the interpretability is not a concern.
- In such cases, we should use the most flexible model available.


Class 1
Class 2
Non Linear Decision Boundary
Class 1
Class 2
kernel
Class 1
Class 2
Decision surface
Class 1
Class 2
Kernel method

Assessing Model Accuracy

No Free Lunch Theorem

- During this course we would introduce a wide range of machine learning models.
- These models are more complex than the standard linear regression approach.
- The question is why do we need so many different machine learning approaches, rather than having a best method?
- In statistics and machine learning, we follow no free lunch theorem.
- For a given data set, one specific approach may give us the best results but some other scientific approach may give better results on a similar but different data set.
- Hence, we need to explore and decide for each data set which approach provides us the best results.
- The most challenging part of the machine learning is to select the approach that can provide us the best results.

Measuring Quality of Fit

- For a given set, we need to evaluate the performance of the machine learning method and measure how well its predictions actually match the observed data.
- We need to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.
- In the regression setting, the most commonly-used measure is the **mean squared error (MSE)**, given by

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

- $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the ith observation.
- The mean square error will be small if the predicted response and the true response are very close.
- It will be large if the predicted response is substantially different from the true response.

Training MSE

- We compute the MSE using the training data that we used to fit the model.
- Hence, we call it training MSE. However, in practice, we are not bothered about the performance of the model on the training data.
- Rather, we are interested in the accuracy of prediction that we get using the previously unseen test data.
- The question arises, why we are interested in unseen test data not in training data?
- Suppose our goal is to develop a machine learning model to predict the stock price base on historical stock returns.
- We can use the last 6 months stock return data to train our model.
- We would not be interested in how well the model is predicting the stock price for a past date.
- Rather we would be interested in how well the model can predict the stock price the next day or the next month.

Training MSE

- Similarly, if we have clinical data that includes weight, blood pressure, height, age, and family history of disease for a number of patients.
- We also have information about whether each patient has diabetes.
- This data can be used to train a machine learning model to predict the risk of diabetes based on clinical observations.
- In practice, we are interested accurately predicting diabetes risk for future patients based on their clinical observations.
- We do not want to know how accurately the model predicts diabetes risk for patients used to train the model.
- We already know which of those patients have diabetes.

Test MSE

- Mathematically, suppose we fit our machine learning model on the training measurements $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ to estimate $\hat{f}$.
- Using the machine learning model, we can compute $\hat{f}(x_1)$, $\hat{f}(x_2), \ldots, \hat{f}(x_n)$.
- Suppose these outputs are approximately equal to $y_1, y_2, \ldots, y_n$, the training MSE would be small.
- But we are not interested in whether $\hat{f}(x_i) \approx y_i$.
- We are interested to know whether $\hat{f}(x_0) \approx y_0$ where (x_0, y_0) is a previously unseen test measurement that was not used to train the machine learning model.
- In this case, we are interested in choosing the model that gives the lowest test MSE not the lowest training MSE.

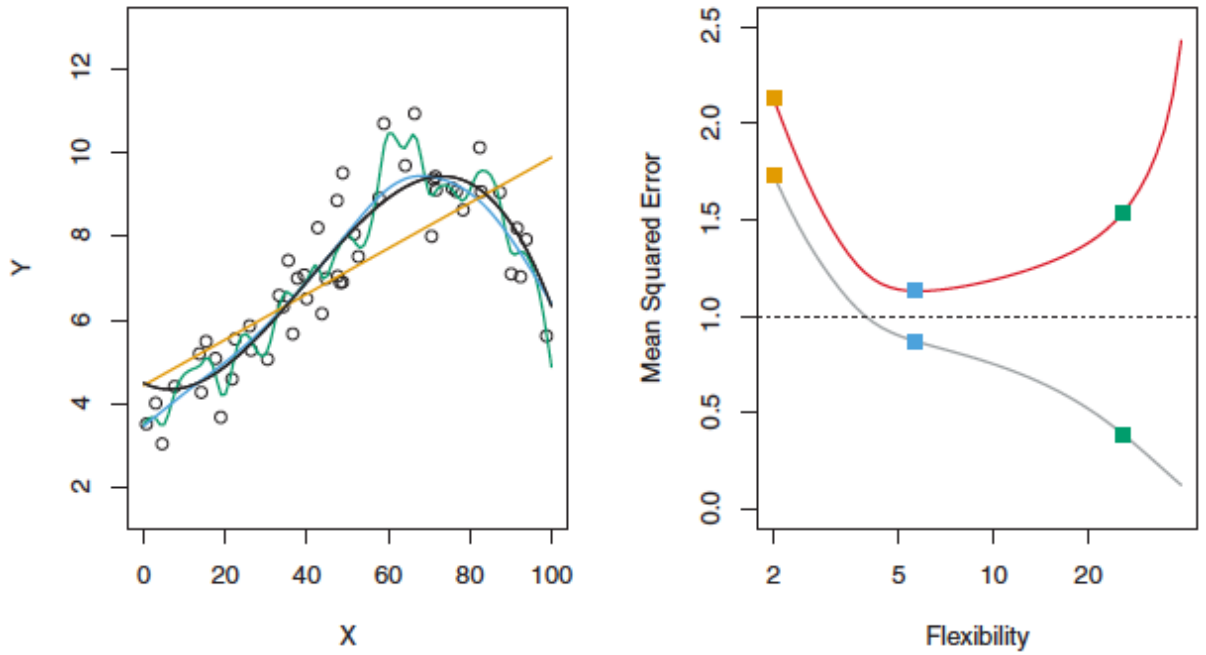
Model Selection

- How do we select a model that results in the minimization of the MSE?
- In certain situations, the test data set might be available.
- In other words, we have a set of observations that we did not use to train the machine learning method.
- In this case, we can evaluate the test observations and select the model with the smallest test MSE.

Model Selection

- On the other hand, in certain situations, the test observations are not available.
- In such situations, we can select the model with the smallest training MSE.
- Even though the training MSE and test MSE appear to be closely related, there is no guarantee that the model with the lowest training MSE will also have the lowest test MSE.
- For many machine learning methods, the training set MSE can be quite small, but the test MSE is often much larger.

Model Selection



Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE overall methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

Model Selection

- The figure above shows the phenomenon using a simple example.
- The left-hand side panel shows the true f by the black curve.
- The three possible estimates for f that have been obtained using methods with increasing levels of flexibility have been shown in orange, blue, and green curves.
- The orange line is the linear regression fit, which is relatively inflexible.
- The blue and green curves were produced using smoothing splines with different levels of smoothness.
- We can see that as the level of flexibility increases, the curves fit the observed data more closely.
- The green curve is the most flexible and matches the data very well but it fits the true f (shown in black) poorly because it is too wiggly.
- We can adjust the level of flexibility of the smoothing spline fit to produce different fits to the data.

Model Selection

- Now we will analyze the right-hand panel.
- The grey curve shows the average training MSE as a function of flexibility. We can also refer to the number of smoothing spines as the degrees of freedom.
- The degrees of freedom summarizes the flexibility of a curve.
- The orange, blue, and green squares represent the MSEs corresponding to the curves in the left-hand panel.
- A more restricted and hence smoother curve such as linear regression has fewer degrees of freedom than a wiggly curve.
- The linear regression is at the most restrictive end, with two degrees of freedom. The training MSE declines monotonically as flexibility increases.
- In this example the true f is non-linear, and so the orange linear fit is not flexible enough to estimate f well.
- The green curve has the lowest training MSE of all three methods since it corresponds to the most flexible of the three curves fit in the left-hand panel.

Model Selection

- We have demonstrated the test MSE using the red curve in the right-hand panel.
- The test MSE along with training MSE initially decline with the increase in the level of flexibility.
- At a certain point the test MSE levels off and then it starts to increase again.

Model Selection

- On the other hand, the orange and green curves both have high test MSE.
- The blue curve minimizes the test MSE because the blue curve appears to estimate f, the best among all the curves.
- The horizontal dashed line represents the irreducible error, $Var(\epsilon)$ that is the lowest achievable test MSE among all possible methods.
- Hence, the smoothing spline represented by the blue curve is close to optimal.

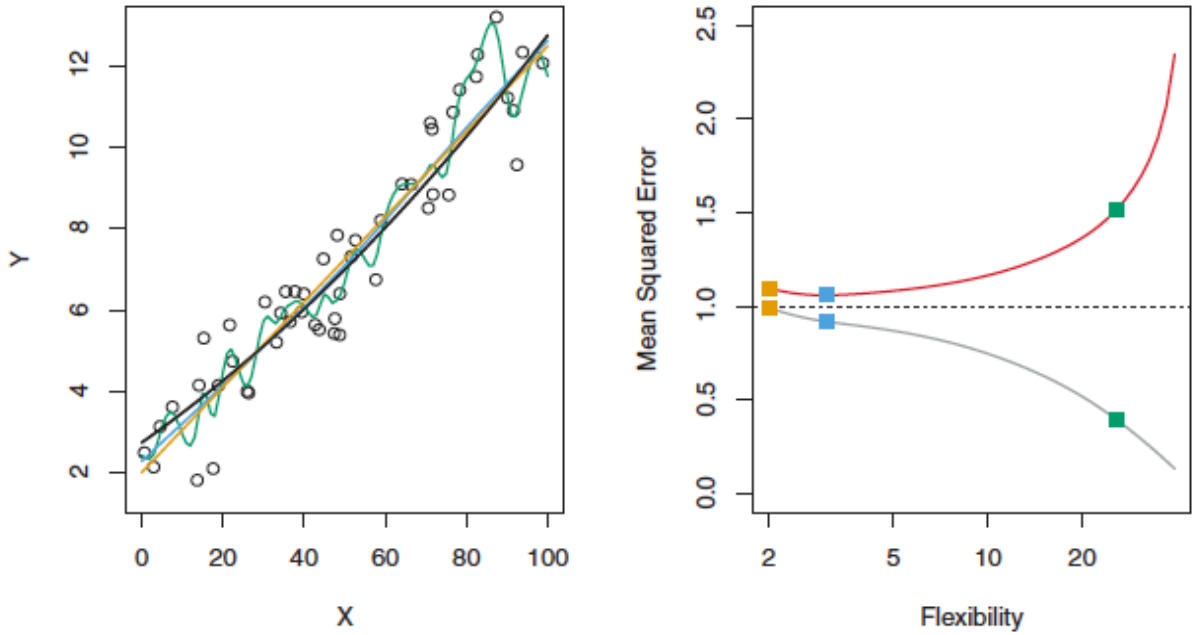
Model Selection

- From the right-hand side panel, we can see that as the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and a U-shape in the test MSE.
- This is a fundamental property of machine learning that holds regardless of the particular data set at hand and regardless of the statistical method being used.
- As model flexibility increases, training MSE will decrease, but the test MSE may not.
- When a given method yields a small training MSE but a large test MSE, we are said to be overfitting the data.
- This happens because our machine learning procedure is working too hard to find patterns in the training data, and maybe picking up some patterns that are just caused by random chance rather than by true properties of the unknown function f.

Model Selection

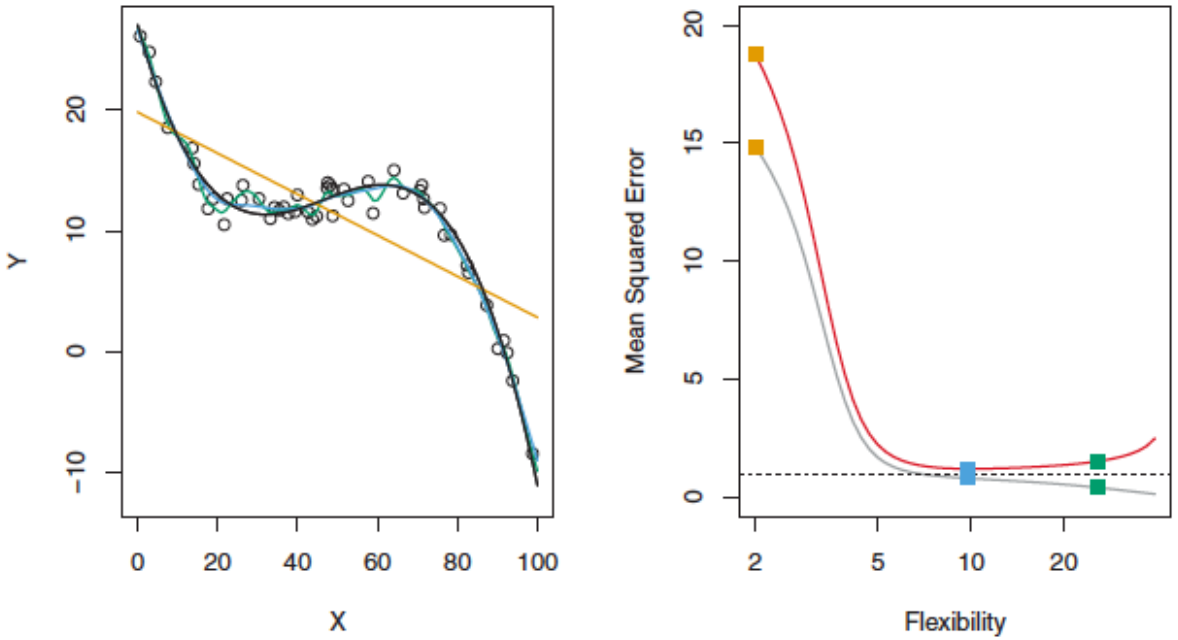
- When we overfit the training data, the test MSE will be very large because the supposed patterns that the method found in the training data simply don’t exist in the test data.
- Note that regardless of whether or not overfitting has occurred, we almost always expect the training MSE to be smaller than the test MSE because most machine learning methods either directly or indirectly seek to minimize the training MSE.
- Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE.

Model Selection

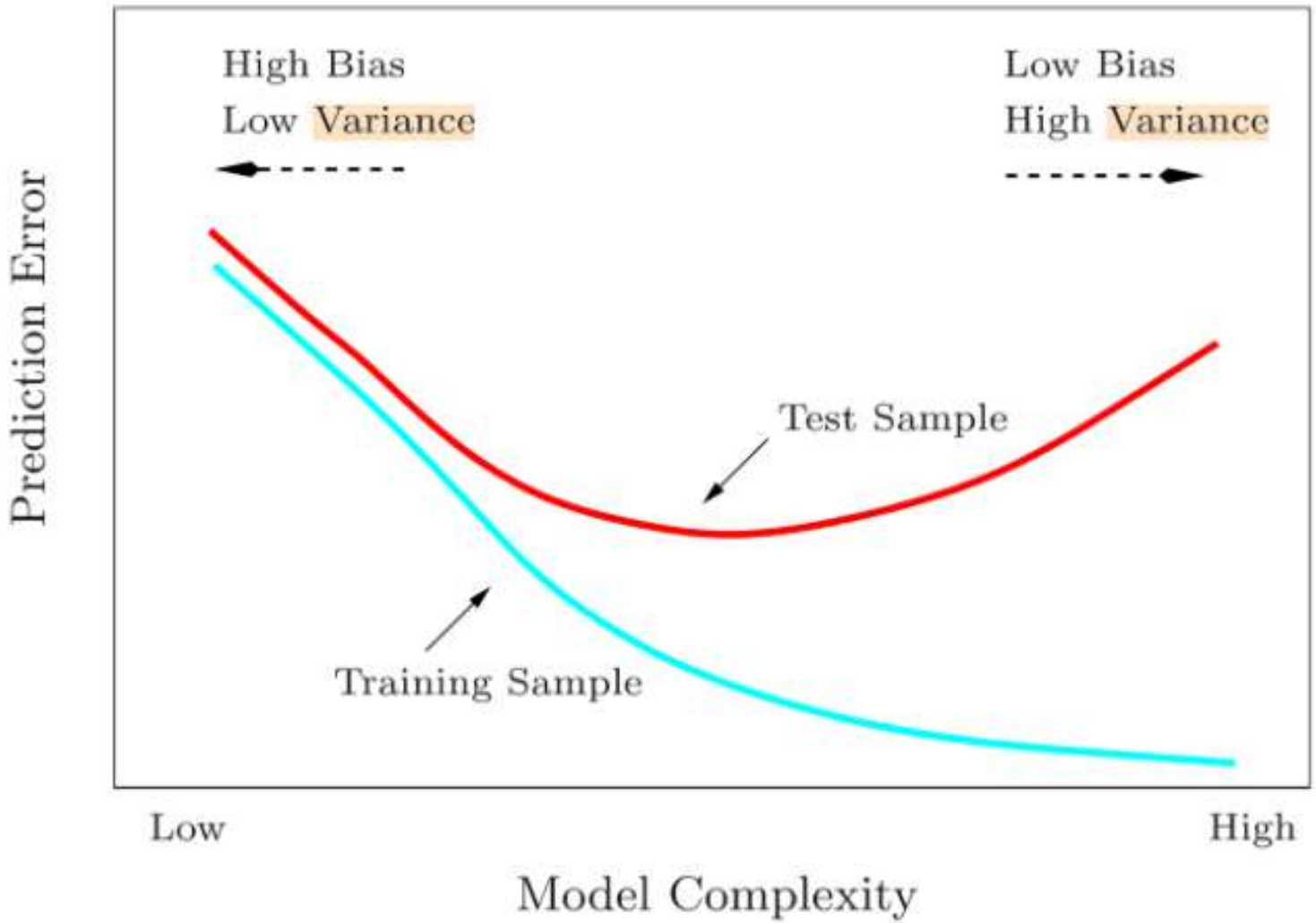


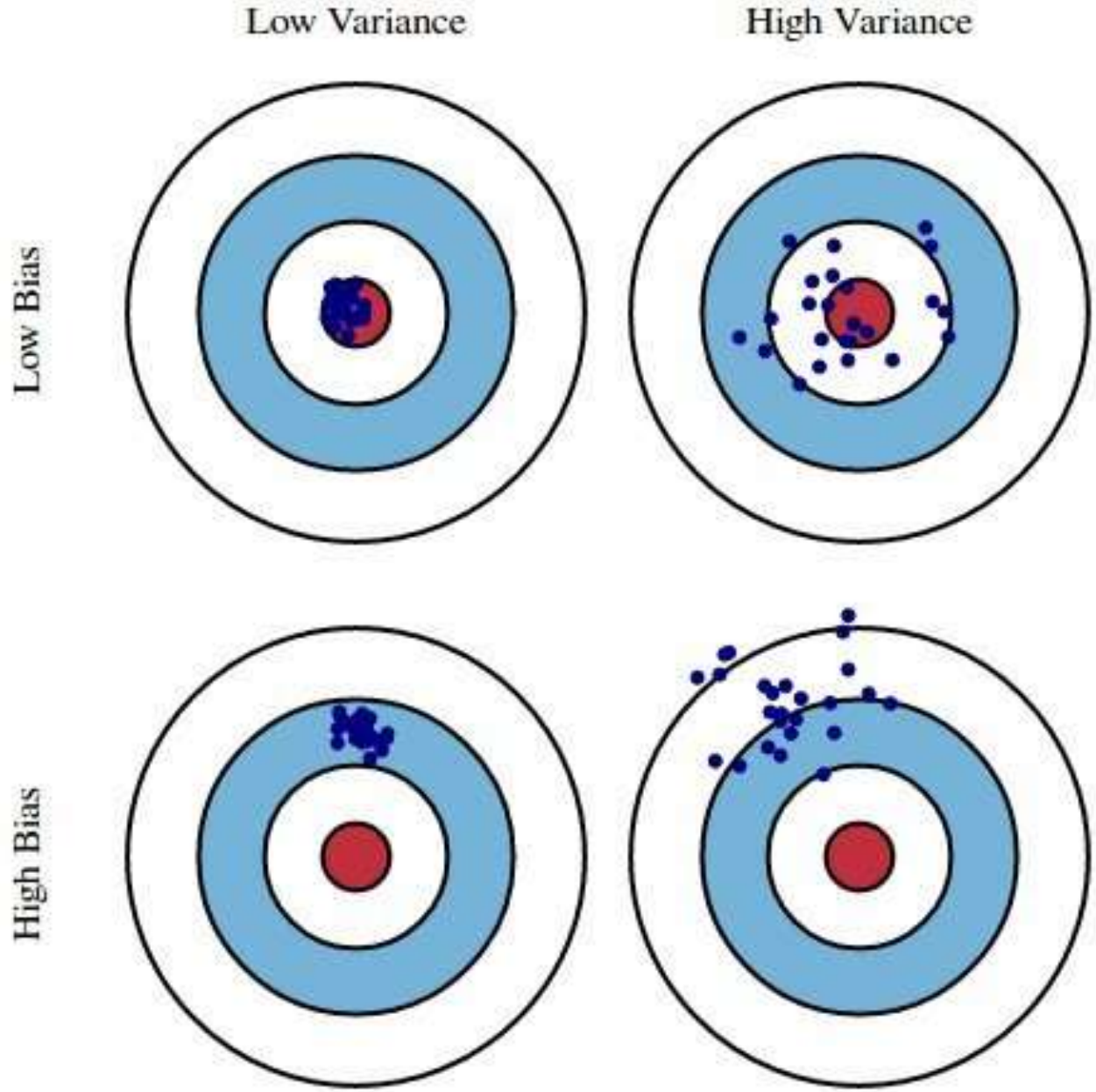
The figure 10 provides another example in which the true f is approximately linear. We can notice that the training MSE decreases monotonically as the model flexibility increases, and that there is a U-shape in the test MSE. However, because the truth is close to linear, the test MSE only decreases slightly before increasing again, so that the orange least squares fit is substantially better than the highly flexible green curve.

Model Selection



The figure 11 displays an example in which f is highly non-linear. The training and test MSE curves still exhibit the same general patterns, but now there is a rapid decrease in both curves before the test MSE starts to increase slowly.


High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Training Sample
Low
High
Model Complexity



Bias – Variance Trade Off

- The U-shape observed in the test MSE curves (the 3 pictures above) turns out to be the result of two competing properties of machine learning methods.
- We can show that the expected test MSE, for a given value x_0, can be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$, and the variance of the error terms ϵ

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\epsilon)$$

Bias – Variance Trade Off

- The notation $E\left(y_0 - \hat{f}(x_0)\right)^2$ signifies that **expected test MSE**.
- It refers to the average test MSE that we can obtain by repeatedly estimating f using a large number of training sets and by testing each at x_0.
- We can compute the expected test MSE by averaging $E\left(y_0 -$

Bias – Variance Trade Off

- From the equation above, we can infer that in order to minimize the expected test error, we should select the method, that can achieve the low variance and low bias simultaneously.
- We can see that the variance is a nonnegative quantity and squared bias is also a nonnegative quantity.
- So, it is not possible to achieve the expected test MSE below $Var(\epsilon)$, the irreducible error.

Meaning of Bias – Variance Trade Off

- The amount by which $\hat{f}$ would change if we use a different training data set to estimate it is **variance**.
- Because we use the training data to fit the machine learning models, we would get different $\hat{f}$ by using different training data sets.
- However, ideally the $\hat{f}$ should not vary too much if we use different training data sets.
- But, due to the high variance of the model, small changes in the training data set can lead to large changes in $\hat{f}$.
- In general, we observe that more flexible machine learning models have higher variance.

Meaning of Bias – Variance Trade Off

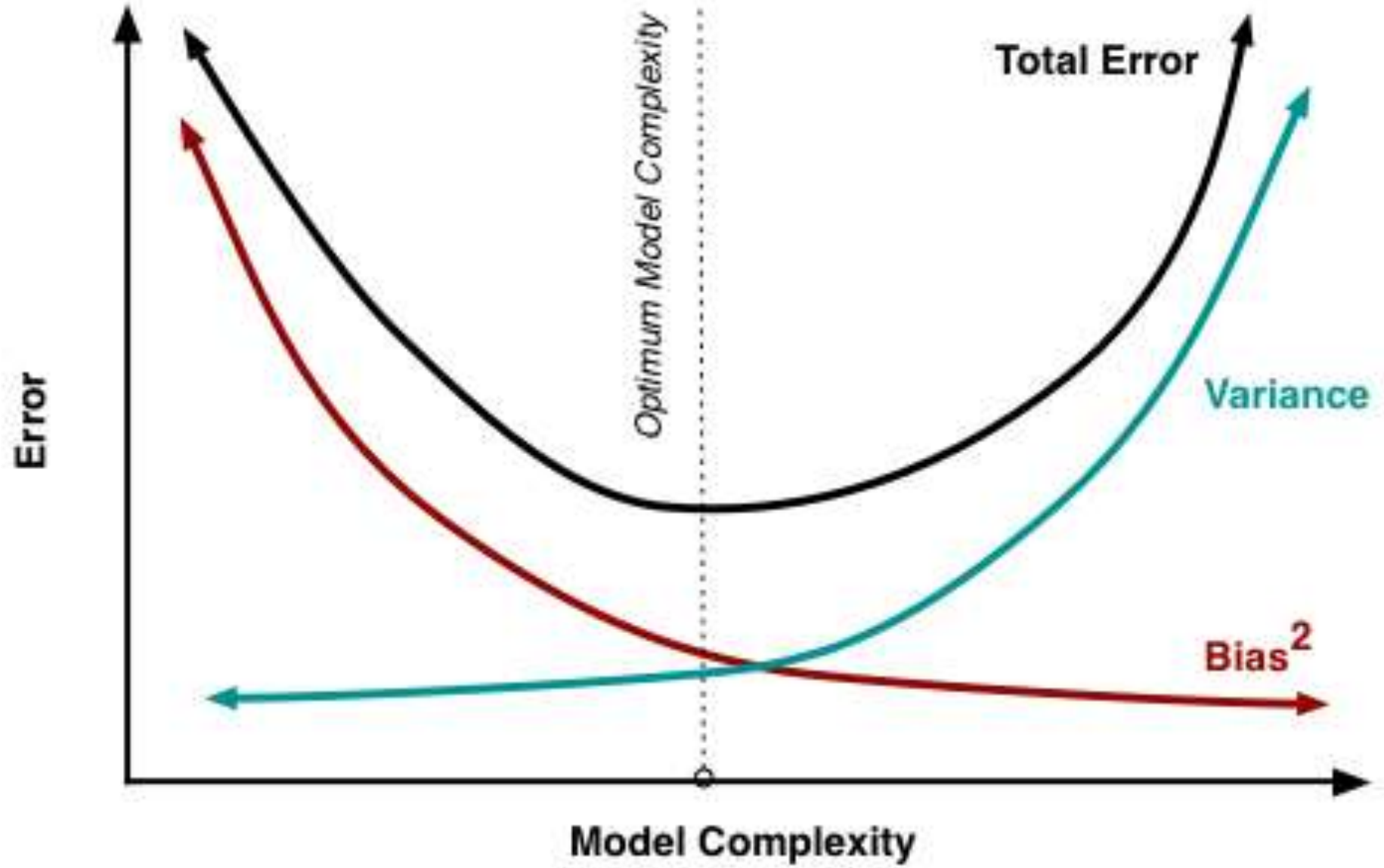
- Let's consider the figure 10 again.
- The flexible green curve is very close to the observations.
- It has high variance.
- Because if we change any one of these data points, $\hat{f}$ may change considerably.
- On the other hand, the orange least squares line is relatively inflexible and has low variance.
- Because if we move any single observation, there will only be a small shift in the position of the line.

Meaning of Bias – Variance Trade Off

- **Bias** refers to the error that is we get by approximating a real-life and extremely complicated problem using a much simpler model.
- For example, the assumption in the case of linear regression is that there is a linear relationship between Y and $X_1, X_2, \ldots, X_p$.
- It is unlikely for a real-life problem to have such a simple linear relationship.
- Hence in this case, performing linear regression will certainly result in some bias in the estimate of f.
- In Figure 11, we can see that the true f is non-linear. Irrespective of number of training observations that we give, we cannot produce an accurate estimate using linear regression.

Meaning of Bias – Variance Trade Off

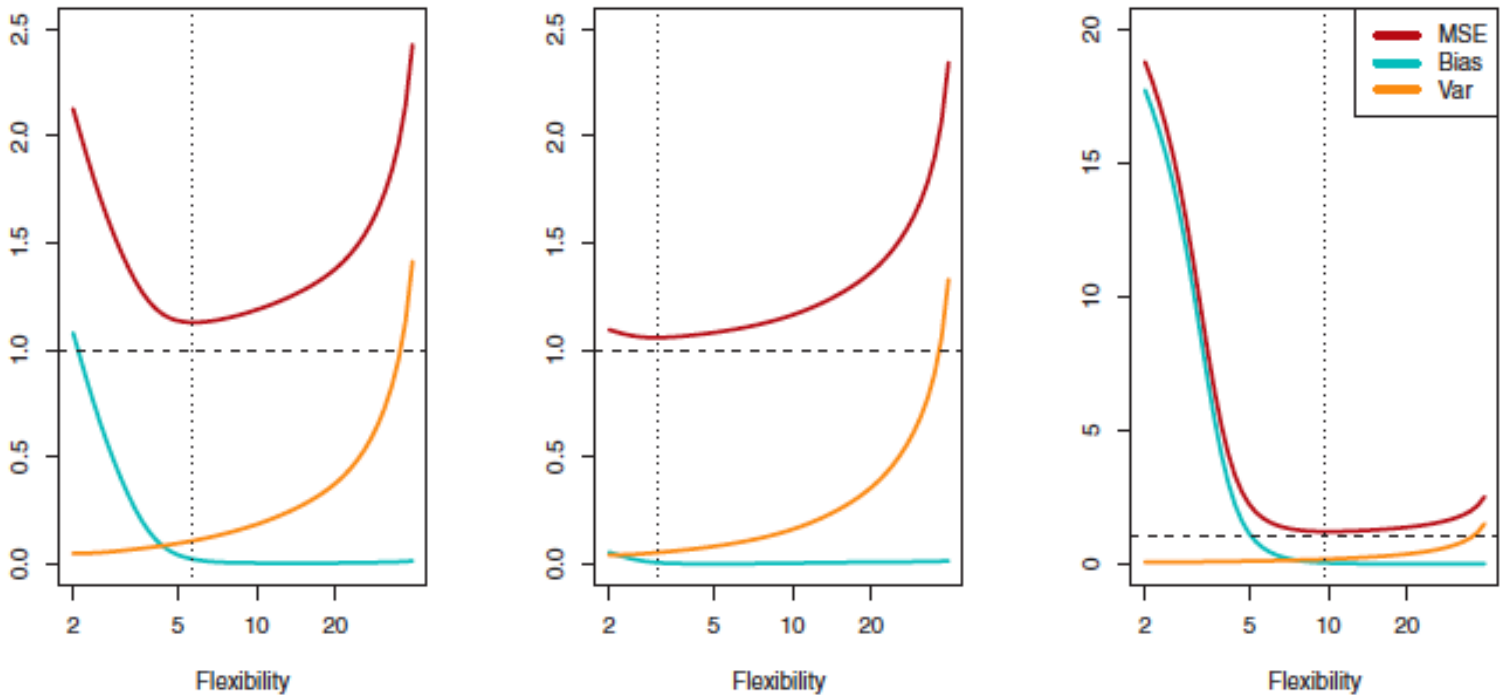
- Hence, in this example, linear regression results in high bias.
- However, in Figure 10, the true f is very close to linear.
- So, if we provide enough data, linear regression should be able to produce an accurate estimate.
- Generally, more flexible methods result in less bias.


Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity

Meaning of Bias – Variance Trade Off

- We can generalize the concept. As the model becomes more flexible, the variance increases and the bias decreases.
- By analyzing the relative rate of change of these two quantities, we can determine whether the test MSE will increase or decrease.
- As the flexibility of the model increases, the bias tends to initially decrease faster than the variance increases.
- As a result, the expected test MSE decreases.
- After some point an increase in flexibility has little impact on the bias but it starts to significantly increase the variance.
- Due to this, the test MSE increases.
- You can note this pattern of decreasing test MSE followed by increasing test MSE in the right-hand panels of Figures 9–11.

Meaning of Bias – Variance Trade Off



The three plots in Figure 12 illustrate relationship between bias and variance for the examples in Figures 9–11.

Meaning of Bias – Variance Trade Off

- In each case the blue solid curve represents the squared bias, for different levels of flexibility, while the orange curve corresponds to the variance.
- The horizontal dashed line represents $Var(\epsilon)$, the irreducible error. Finally, the red curve that corresponds to the test set MSE, is the sum of these three quantities.
- In all three cases, the variance increases and the bias decreases as the method's flexibility increases.
- However, the flexibility level corresponding to the optimal test MSE differs considerably among the three data sets, because the squared bias and variance change at different rates in each of the data sets.
- In the left-hand panel of Figure 12, the bias initially decreases rapidly, resulting in an initial sharp decrease in the expected test MSE.

Meaning of Bias – Variance Trade Off

- On the other hand, in the center panel of Figure 12 the true f is close to linear, so there is only a small decrease in bias as flexibility increases, and the test MSE only declines slightly before increasing rapidly as the variance increases.
- Finally, in the right-hand panel of Figure 12, as flexibility increases, there is a dramatic decline in bias because the true f is very non-linear.
- There is also very little increase in variance as flexibility increases.
- Consequently, the test MSE declines substantially before experiencing a small increase as model flexibility increases.

Meaning of Bias – Variance Trade Off

- The relationship between bias, variance, and test set MSE in Figure 12 is referred to as the **bias-variance trade-off**.
- Good test set performance of a machine learning method requires low variance as well as low squared bias.
- This is referred to as a trade-off because it is easy to obtain a method with extremely low bias but high variance (for instance, by drawing a curve that passes through every single training observation) or a method with very low variance but high bias (by fitting a horizontal line to the data).
- The challenge lies in finding a method for which both the variance and the squared bias are low.

Thanks

Samatrix Consulting Pvt Ltd

Samatrix.io